AUDYOGIK SHIKSHAN MANDAL
Since 1983
ASM
Education • Research • Consultancy

ASM Group of Institutes

incon -VII 2012

International Conference on Ongoing Research in Management and IT

In Association With

AMMI

CETYS UNIVERSIDAD Mexico

Conference Proceeding

ISBN 978-81-921445-1-1

3rd and 4th March 2012
ASM Campus Pimpri,
Pune, India.

VENUE :
ASM Group of Institutes, S.No. 29/1+2A,
CTS No. 4695, Opp. PCMC Building Near
Sterling Honda Showroom, Old Pune Mumbai
Highway, Pimpri, Pune 410018.
Ph: +91-20-66351700, 27475090, 27478666.
Email: ibmrc@vsnl.in
Website: www.asmedu.org

ASM Group of Institutes : IBMR | IIBR | IPS | ICS | IMCOST | CSIT

RESEARCH IN MANAGEMENT AND IT

This E-Book Is Collection Of The Papers Submitted By Various Authors To INCON VII 2012 ***" International Conference On Ongoing Research In Management And It"*** Organized By ASM Group Of Institutes In Association With AMMI And CETYS Universidad Mexico

Conference proceedings are published on conference date, having ISBN No. 978-81-921445-1-1

Disclaimer : The views expressed in the research papers are of the authors. ASM Group OF Institutes does not hold any liability for the contents of papers.

ABOUT ASM (AUDYOGIK SHIKSHAN MANDAL)

For over 28 years, learning and research at ASM Group of Institutes has played a major part in grooming students from across the country and abroad for successful careers in the industry. From 1983 to the present day, we've been leading the way in research, consultancy and education, making ground-breaking progress in areas that span industries such as Engineering, IT, FMCG, Communications, etc.

Excellent sporting facilities, a busy calendar of social events and a global reputation for teaching makes ASM Group of Institutes a great place to study and we're proud of the calibre of our academic community. Our vision is to provide the best education to each and every student who joins any of our colleges.

Each year, over 3,000 students from across India & Abroad study at the ASM Group of Institutes, enhancing our reputation as a truly Global Institute. We are committed to enabling all our students and staff to profit from a culture of learning, aligned with our research & development ethos. The campuses are safe and friendly, with an impressive set up of state-of-the-art computer labs and wi-fi enabled hot spots and modern buildings.

ABOUT CETYS UNIVERSITY

CETYS University is an educational institution of excellence, situated in the state of Baja California, Mexico. The CETYS University seeks, just what it establishes in its mission, "To contribute to the formation of people with the moral capacity and necessary intellectual to participate in important form in the cultural, social, and economic improvement of the country".

ABOUT AMMI

The Association of Management of MBA/MMS Institutes in Maharashtra is registered as "Association of Management of MBA/MMS Institutes "bearing No. MAH/695/2005 Pune dated 26/05/2005.

The chief role of the Association is to be a significant contributor to the field of management and to revolutionizing the study of management at the post graduate level.

ASM INCON 2012

INTERNATIONAL CONFERENCE 2012 is the 7th in the chain of Conferences organized by ASM Group of Institutes in Pune (India).

The main theme for this Conference is " Ongoing Research and Emerging Perspectives in the fields of Management Studies and Information Technology".

This theme is further broken up into four broad heads namely: Finance, Marketing, Human Resource Management and Information Technology.

Each one of the heads will form the domain of a separate group of experts who have further disaggregated these heads into sub themes as indicated towards the end. Research papers for purposes of discussions have been solicited from the leading experts in the line for deliberations at the Conference.

The International Conference organized so far has been extremely useful not only for the researchers in the academic circles but also has been offering an excellent platform for interactions with the Academicians and Representatives of the Corporate sector as well as the Government. As a matter of fact the conference has been serving as a tool for the decision makers in the fields of Academics , Corporate houses, Government undertakings and the Non- Government undertakings to come together and solicit solutions to the problems confronting them from time to time.

STEERING COMMITTEE

- Dr. Asha Pachpande - Secretary, ASM
- Dr. Sandeep Pachpande - Chairman, ASM
- Dr. Santosh Dastane - Director Research, ASM's IBMR
- Prof. S. D. Mathur - Director General, ASM's IIBR
- Dr. G. B. Patil - Director, ASM's IPS
- Prof. Ashish Dixit - Director, ASM's ICS
- Dr. Prakash Deshpande - Director, ASM's IIBR
- Dr. Carlos Rodriguez - CETYS, Mexico
- Dr. Scott Venezia - Head of Business & Mgmt division, CETYS, Mexico
- Dr. Patricia Valdez - CETYS, Mexico
- Dean Jorge Sosa - Dean, School of Engineering, Maxicali Campus, CETYS Mexico
- Dr. Gopalchandra Banik - Director, ASM's IMCOST, Mumbai
- Dr. V. P. Pawar - Director, ASM's IBMR, MCA
- Dr. S. P. Kalyankar - Professor, ASM's IBMR

ADVISORY COMMITTEE

- Dr. Ashok Joshi - Dean, Faculty of Management, University of Pune
- Dr. Vinayak Deshpande - Prof. & Head, Department of Management, Nagpur University
- Capt. Dr. C.M. Chitale - Prof. & Head, MBA, University of Pune
- Dr. Anil Keskar - Dean (Academics), Symbiosis University
- Dr. S. B. Kolte - Director, IMSCDR, Ahmednagar
- Dr. Sharad Joshi - Management Consultant
- Dr. Kaptan Sanjay - Prof. & Head, Depar tment of Commerce,
- Dr. S. G. Bapat - Director, S G Bapat & Associates
- Dr. D. D. Balsaraf - Principal, Indrayani Mahavidyalay Talegaon, Pune
- Dr. Sanjay R. Mali - Principal, Arts & Commerce College, Lonavala
- Dr. E. B. Khedkar - Director, Allard Institute of Management
- Dr. D. Y. Patil - Director Bharati Vidyapeeth, Mumbai
- Dr. G. K. Shirude - Director, Prin. N. G. Naralkar Institute
- Dr. Nitin Ghorapade - Principal, Prof. Ramkrishna More College
- Dr. T. Shivare - Principal, Hinduja College, Mumbai

In the global business environment, business cycles are regular phenomena. To minimize the impact of crests and troughs of cycles, emerging trends and ongoing research should provide shock absorber. INCON VII-2012, primarily intends to focus on the aspect of emerging trends in global business environment.

Thus, on-going research in Management and IT is the focused area of ASM's INCON VIIth Edition

INCON VII Opening and Valedictory Function
Guests : Dean Jorge Sosa, Dr. E.B. Khedkar, Dr. Ashok Joshi, Dr Asha Pachpande, Dr. Sandeep Pachpande, Dr. Santosh Dastane

INDEX

016	Systems	Mr. P.D. Joshi	
IT 017	Feature Based Audio Segmentation Using K-Nn And Gmm Methods	Mrs.Borawake Madhuri Pravin Prof.Kawitkar Rameshwar Prof.Khadtare Mahesh	166
IT 018	Governance Of IT In Indian Banks – An Immense Need	Mrs. K. S. Pawar Dr. R. D. Kumbhar	180
IT 019	Tangible User Interface:Unifying The New Genration Of Interaction Styles	Yogesh H.Raje Dr.Amol C. Goje	198
IT 020	Impact Of Information Technology On The Development Of Rural Economy Of India	Yogesh Raaje Abhay Chounde	208
IT 021	Distributed Virtual Compiler Editor	Sameeran J. Tammewar	215
IT 022	Challenges In Internet Traffic Management	Mrs. Minaxi Doorwar Ms. Archana Walunj	226
IT 023	System Dedicated To Process Singing Voice For Music Retrieval In Indian Classical Terminology	Jui Jamsandekar Mayuri Karle	235
IT 024	Audio Segmentation	Mrs. Borawake Madhuri Pravin Prof. Kawitkar ameshwar Prof. Khadhatre Mahesh	246
IT 025	Linguistic Analysis To Detect Ambiguities And Inaccuracies In Use-Cases	Saket Guntoorkar, Rohan Kokandakar Saiprasad Dhumal, Yogesh Arvi	261
IT 026	An External Storage Support For Mobile Application With Scarce Resources	Pratima.J.Shedge, Nishad Rajwade	269
IT 027	Best Implementations And High-Performance Storage Virtualization Architecture	Sachin V. Joshi Miss. Shrutika A Hajare Devendra R. Bandbuche	286
IT 028	Recognition Of Guilty Agent In Information Leakage	Poonam Sahoo Kanchan Garud Sarika Kate Sharayu Padmane	297
IT 029	Instant Academic Notices On Phones	Satish Kamble Prachi Tongaonkar Remya Mogayan Shreya Kadam	303
IT 030	Web Search Optimization By Mining Query Logs	Ashish Pathak Sonam Jain	310

		Priyanka Chavan Shashikant Gujrati	
IT 031	Image Identification Using CBIR	Suvarna Vitthal Khandagale Annu Shabbir Tamboli Kamthe Sweety Haridas Salunke Rajashri R	319
IT 032	Performance Analysis Of Reactive Routing Protocols For Mobile Ad Hoc Networks	Deepika A.Sarwate Sheetal S.Patil	333
IT 033	A Comparative Study Of Selected Security Tools For Personal Computers	Deepika A.Sarwate Sheetal S.Patil	353
IT 034	Application Of Cloud Computing On E-Learning	Sachin V. Joshi Shrutika V Hazare Devendra R. Bandbuche	371
IT 035	Secure Mobile Ad Hoc Network	Varsha S. Upare	379
IT 036	Web Service Testing Tool	Rohit Kishor Kapadne Ishwar M. Mali Mangesh Gailkwad Sunny Kawade	388
IT 037	"An Approach to solve system Dynamics Problems using Fuzzy Logic"	Mrs. Aparna B Kodgirwar Mrs. Sheetal Deshmukh	407
IT 038	Comparison of Various Filters Applied on Different Types of Noises in Images under Image Processing Techniques	Divya Khandelwal Pallavi Gillurkar	419
IT 039	Spatial Data Mining in Distributed DBMS	Prof. Hidayatulla K. Pirjade	429
IT 040	Enhance the privacy issues related to SQL Azure	Mrs.Priya Joshi	435
IT 041	Free/Open Source Software:Opportunities And Challenges For Modeling And Simulating	Mrs. Ashwini R. Bhirud	446
IT 042	Green IT – With Emerging Approaches	Prof. Asheesh Dixit Santosh. B. Potadar Yogeshkumar. V. Gaikwad	456
IT	A Look into Density Based	Ms. Roopa Praveen	466

IT 001

Fuzzy Computational intelligence for knowledge discovery in databases

Mr. Gajanan. M. Walunjkar

Asst Prof. Dept of IT,

Army Institute of Technology,Pune-15,University of pune

Maharashtra ,India

gmgaju@yahoo.co.in

I. ABSTRACT

The aim of the Steps of Knowledge Discovery domain is to show how the elements of CI (Computational intelligence)can be used in data mining and how fuzzy information processing can be situated within this general and comprehensive process. In the remaining sections, basic definitions, widely applied methods and tools for clustering (Classical Fuzzy Cluster Analysis section), visualization (Visualization of High Dimensional Data section), classification (Fuzzy Classifier Systems for Effective Model Representation section), and association rule mining (Fuzzy Association Rule Mining section) are discussed, including the related knowledge representation, identification, and reduction methods.

The aim is to give an overview about fuzzy data mining methods. It should be emphasized that this title is equivocal in some sense because it has two meanings. Fuzzy data mining methods can mean *data mining methods* that are fuzzy methods as well; on the other hand, it can also mean approaches to analyze *fuzzy data*. In some sense, the later ones are fuzzy methods as well but the conceptions are different. Fuzzy data mean imprecise, vague, uncertain, ambiguous, inconsistent, and/or incomplete data. Therefore, the source of uncertainty is the data themselves. It is very important to develop methods that are able to handle this kind of data because data from several

information sources might be fuzzy (e.g., from human expert who describes their knowledge in natural language).

This paper deals with data mining methods based on fuzzy techniques. The data are crisp and can be given in absolute ($N \times n$ matrix) or relative form ($N \times N$ matrix, where N and n are the numbers of samples and attributes, respectively. By absolute data, the values of the attributes are given. "Relative data means that the data's values are not known, but their pairwise distance is known. The approaches the data are analyzed with handle the uncertainty on the basis of fuzzy logic

Keyword: KDD(Knowledge discovery in database), Fuzzy classifier, data mining.

II. INTRODUCTION

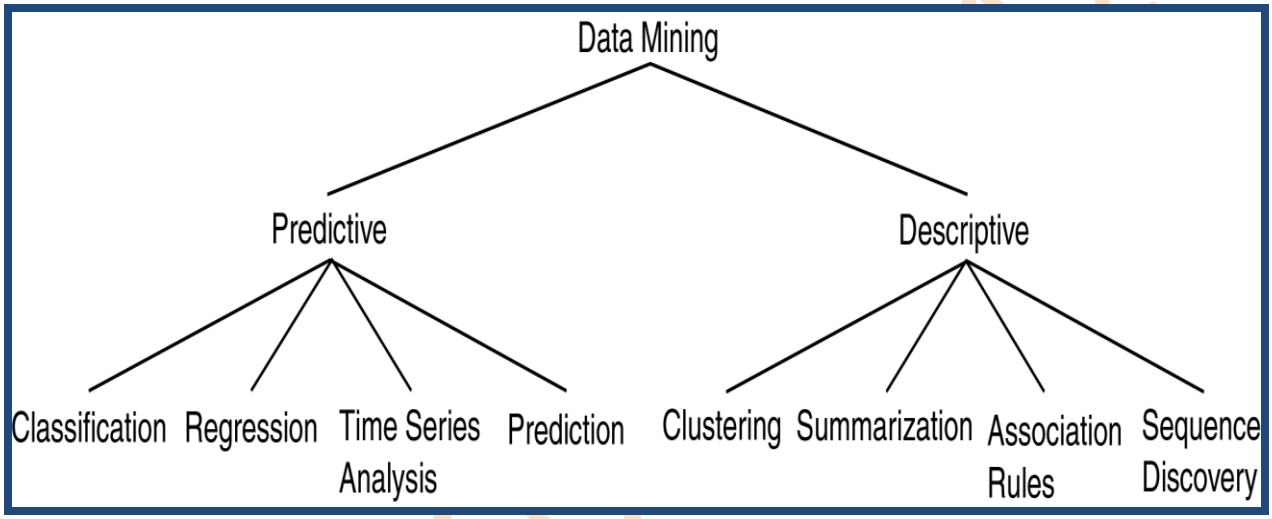
It is a set of Nature-inspired computational methodologies and approaches to address complex problems of the real world applications to which traditional (first principles, probabilistic, black-box, etc.) methodologies and approaches are ineffective or infeasible. It primarily includes Fuzzy logic systems, Neural Networks and Evolutionary Computation. In addition, CI also embraces techniques that stem from the above three or gravitate around one or more of them, such as Swarm intelligence and Artificial immune systems which can be seen as a part of Evolutionary Computation; Dempster-Shafer theory, Chaos theory and Multi-valued logic which can be seen as off-springs of Fuzzy Logic Systems, etc.The characteristic of 'intelligence' is usually attributed to humans. More recently, many products and items also claim to be 'intelligent'. Intelligence is directly linked to the reasoning and decision making.

Data mining is the central step in a process called knowledge discovery in databases, namely the step in which modeling techniques are applied. Several research areas like statistics, artificial intelligence, machine learning, and soft computing have contributed to its arsenal of methods. In this paper, however, we focus on fuzzy methods for rule learning, information fusion, and dependency analysis.[10] In our opinion fuzzy approaches can play an important role in data mining, because they provide comprehensible results (although this goal is often neglected—maybe because it is sometimes hard to achieve with other methods). In addition, the approaches studied in data mining have mainly been oriented at highly structured and precise data. However, we expect that the analysis of more complex heterogeneous information source like texts, images, rule bases etc. will become more important in the near future.[11] Therefore we give an outlook on information

mining, which we see as an extension of data mining to treat complex heterogeneous information sources, and argue that fuzzy systems are useful in meeting the challenges of information mining. Knowledge discovery in databases (KDD) is a research area that considers the analysis of large databases in order to identify valid, useful, meaningful, unknown, and unexpected relationships.

III. FUZZY DATA MINING TASKS

Fig 1 : Fuzzy data mining tasks



There is another important fact that links KDD and fuzzy sets: in many cases data is inherently imprecise or uncertain, and several fuzzy relational, deductive and object-oriented database models have been developed in order to cope with this. A more usual case, that provides a similar scenario, is that of fuzzy data obtained from crisp data in the preprocessing step by aggregation, summarization or change of granularity level. The analysis of such information requires the development of specific tools as fuzzy extensions of existing ones.

Fuzzy Data mining involves following classes of tasks

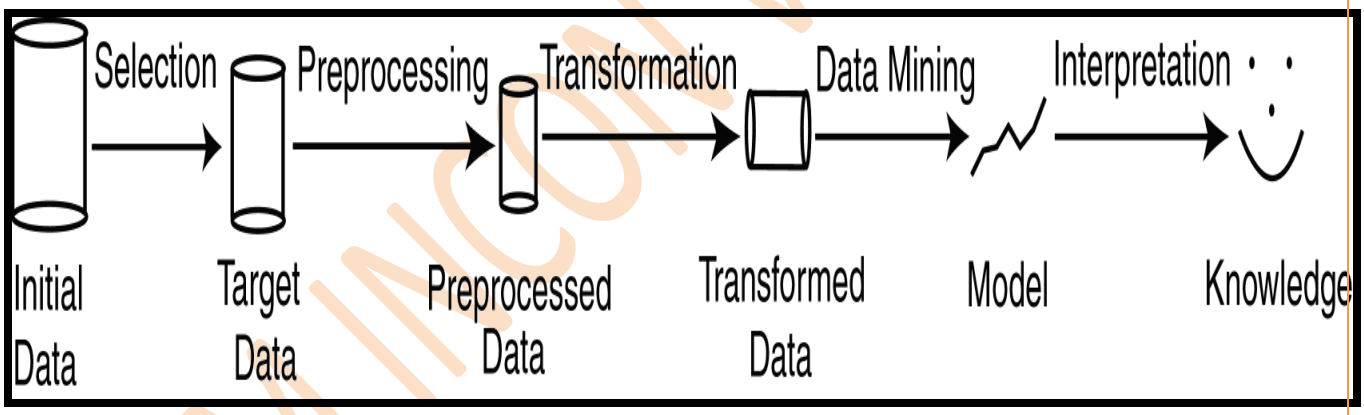
- Classification maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- Regression is used to map a data item to a real valued prediction variable.
- Clustering groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation

– Partitioning
■ Summarization maps data into subsets with associated simple descriptions.
– Characterization
– Generalization
■ Link Analysis uncovers relationships among data.
– Affinity Analysis
– Association Rules
– Sequential Analysis determines sequential patterns.

Fuzzy Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

IV. KNOWLEDGE DISCOVERY IN DATABASES

Fig. 2 : KDD Process



■ Selection: Obtain data from various sources.
■ Preprocessing: Cleanse data.
■ Transformation: Convert to common format. Transform to new format.
■ Data Mining: Obtain desired results.
Interpretation/Evaluation: Present results to user in meaningful manner.

Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data.

Data Mining: Use of algorithms to extract the information and patterns derived by the KDD process

FCM ALGORITHM:

Input: n data objects, number of clusters

Output: membership value of each object in each cluster

Algorithm:

1. Select the initial location for the cluster centres

2. Generate a new partition of the data by assigning each data point to its closest centre.

3. Calculate the membership value of each object in each cluster.

4. Calculate new cluster centers as the centroids of the clusters.

5. If the cluster partition is stable then stop, otherwise go to step2 above.[6,7]

V. DATA MINING & CLUSTERING

Data mining or knowledge discovery in databases (KDD) is the process of discovering useful knowledge from large amount of data stored in databases, data warehouses, or other information repositories.[1] Data mining is a hybrid disciplinary that integrates technologies of databases, statistics, artificial intelligent. Recently, a number of data mining applications and prototypes have been developed for a variety of domains, including marketing, banking, finance, manufacturing, and health care other types of scientific data.[2, 3] The more common model functions in data mining include Classification, Clustering, Discovering association rules, Summarization, Dependency modeling and Sequence analysis. [3] Soft computing methodologies like fuzzy sets, neural networks, and genetic algorithms are most widely applied in the data mining. Fuzzy sets copes with uncertainty in data mining process.

The process of grouping a set of objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. As a branch of statistics, cluster analysis has been extensively studied for many years, focusing mainly on distance-based cluster analysis. Cluster analysis tools based on k-means, k-medoids, and several other methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS (Clementine). In machine learning, clustering is an example of unsupervised learning. In general, the major clustering methods can be classified into some categories: Partitioning methods,Hierarchical methods, Model based method; Grid based method, Clustering high-dimensional data and Constraint based clustering.[1]

VI. APPLICATIONS

The applications are Optimal selection of solvent systems,Roman pottery (terra sigillata), Greek muds and pelloids, Fuzzy system of chemical elements, Romanian and American coals,Intramolecular interactions and catalyst modeling, Assesment of heart disease, Electric network distribution systems.

VII. CONCLUSION

In knowledge discovery and data mining as it is, there is a tendency to focus on purely data-driven approaches in a first step. More model-based approaches are only used in the refinement phases (which in industry are often not necessary, because the first successful approach wins— and the winner takes all). However, to arrive at truly useful results, we must take background knowledge and, in general, non-numeric information into account and we must concentrate on comprehensible models. The complexity of the learning task, obviously, leads to a problem: When learning from information, one must choose between (often quantitative) methods that achieve good performance and (often qualitative) models that explains what is going on to a user. However, in the most successful fuzzy applications in industry such as intelligent control and pattern classification, the introduction of fuzzy sets was motivated by the need for more human-friendly computerized devices that help a user to formulate his knowledge and to clarify, to process, to retrieve, and to exploit the available information in a most simple way. In order to achieve this user-friendliness, often certain (limited) reductions in performance and solution quality are accepted.

So the question is: What is a good solution from the point of view of a user in the field of information mining? Of course, correctness, completeness, and efficiency are important, but in order to manage systems that are more and more complex, there is a constantly growing demand to keep the solutions conceptually simple and understandable. This calls for a formal theory of utility in which the simplicity of a system is taken into account. Unfortunately such a theory is extremely hard to come by, because for complex domains it is difficult to measure the degree of simplicity and it is even more difficult to assess the gain achieved by making a system simpler. Nevertheless, this is a lasting challenge for the fuzzy community to meet.

REFERENCES

[1] J.Allen, A. Christie, W.Fithen, j.McHugh,J.pickel, and E.Stoner, “State of the practice of

Intrusion Detection Technologies”, CMU/SEI-99-TR-028, Carnegie Mellon Software Engg.

Institute. 2000.

[2] KDDCup’1999dataset.http://kdd.ics.uci.edu/databases/kddcup’99/kddcup99.html.

[3] S.theodoridis and K.koutroubas, "pattern Recognition", Academic Press, 1999.

[4] Wikipedia-Cluster Analysis,http://en.wikipedia.org/wiki/cluster_analysis.

[5] Johan Zeb Shah and anomie bt Salim, "Fuzzy clustering algorithms and their application to

chemical datasets", in Proc. Of the post graduate Annual Research seminar 2005, pp.36-40.

[6] Zhengxim Chen, "Data Mining and Uncertain Reasoning-An integrated approach", Willey,

2001.

[7] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds.,Advances in

Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996.

[8] Hussein Aly Abbass, Ruhul Amin, Sarker, Charles S. Newton. Data mining: a heuristic

approach, 2002, Idea Group Publishing.

[9] C. Li and J. Yu, "A Novel Fuzzy C-means Clustering Algorithm," in LNAI 4062, Edited by

G. Wang et al., Springer, pp. 510–515, 2006.

[10] M. Sato-Ilic and L.C. Jain, *Innovations in Fuzzy Clustering*, Springer, 2006.

[11] M. Sato-Ilic, "Weighted Fuzzy Clustering on Subsets of Variables," *International*

Symposium on Signal Processing and its Applications with the International Conference on

Information Sciences, Signal Processing and its Applications, 2007. (Invited paper)

IT 002

Performance Analysis of Reactive Routing Protocols in VANET

Authors:

Poonam Dhamal
Department of IT
GHRCEM COE
Wagholi, Pune, India
poonamdhamal28@gmail.com

Minaxi Rawat
Department of IT
GHRCEM COE
Wagholi, Pune, India
minaxirawat@rediffmail.com

Abstract—VANET (Vehicular Adhoc Network) research field is growing very fast. It has to serves a wide range of applications under different scenario (City, Highway). It has various challenges to adopt the protocols that can serve in different topology and scenario. This paper presents a comparative study of the adhoc routing protocols. The main objective of Vehicular Ad-Hoc Networks is to build a robust network between mobile vehicles so that vehicles can talk to each other for the safety of human beings. VANET hits the protocol's strength due to its highly dynamic features, thus in testing a protocol suitable for VANET implementation we have selected different routing protocols. In this paper, an attempt has been made to compare three well know protocols AODV, AOMDV, DSR by using three performance metrics packet delivery ratio and average end to end delay. The comparison has been done by using simulation tool NS2 which is the main simulator, NAM (Network Animator) and excel graph which is used for preparing the graphs from the trace files.

Keywords-VANET, NAM, AODV, AOMDV, DSR.

Introduction

VANET is a special case of the general MANET to provide communications among nearby vehicles and between vehicles and nearby fixed roadside equipments. VANET networks, nodes are characterized by high dynamic and

mobility, in addition to the high rate of topology changes and density variability [1]. VANETs are a subset of MANETs (Mobile Ad-hoc Networks) in which communication nodes are mainly vehicles. As such, this kind of network should deal with a great number of highly mobile nodes, eventually dispersed in different roads. In VANETs, vehicles can communicate each other (V2V, Vehicle-to-Vehicle communications). They can connect to an infrastructure (V2I, Vehicle-to-Infrastructure) or Infrastructure to Vehicle (I2V) to get some service. This infrastructure is assumed to be located along the roads.

Some motivations of the promising VANET technology include, Increase traveler safety, Enhance traveler mobility, Decrease travelling time, Conserve energy and protect the environment, Magnify transportation system efficiency, Boost on-board luxury but it is not enough many other services can be served by using this technology. The creation of Vehicular Ad Hoc Networks (VANET) has spawn much interest all over the world, in German there is the FleetNet[2] project and in Japan the ITS(Intelligent Transportation System) project. Vehicular ad hoc networks are also known under a number of different terms such as Inter Vehicle communication (IVC), Dedicated Short Range Communication (DSRC) or Wireless Access in Vehicular Environments (WAVE) [3]. The goal of most of these projects is to create new network algorithms or modify the existing for use in a vehicular environment. In the future vehicular ad hoc networks will assist the drivers of vehicles and help to create safer roads by reducing the number of automobile accidents. Vehicles equipped with wireless communication technologies and acting like computer nodes will be on the road soon and this will revolutionize the concept of travelling. VANETs bring lots of possibilities for new range of applications which will not only make the travel safer but fun as well.

Characteristics

VANET has some unique characteristics which make it different from MANET as well as challenging for designing VANET applications.

High Dynamic topology

The speed and choice of path defines the dynamic topology of VANET. If we assume two vehicles moving away from each other with a speed of 60 mph (25m/sec) and if the transmission range is about 250m, then the link between these two vehicles will last for only 5 seconds (250m/ 50ms-1). This defines its highly dynamic topology.

Frequent disconnected Network

The above feature necessitates that in about every 5 seconds or so, the nodes needed another link with nearby vehicle to maintain seamless connectivity. But in case of such failure, particularly in case of low vehicle density zone, frequent disruption of network connectivity will occur. Such problems are at times addressed by road-side deployment of relay nodes.

Mobility Modelling and Prediction

The above features for connectivity therefore needed the knowledge of node positions and their movements which as such is very difficult to predict keeping in view the nature and pattern of movement of each vehicles. Nonetheless, a mobility model and node prediction based on study of predefined roadways model and vehicle speed is of paramount importance for effective network design.

Communication Environment

The mobility model highly varies from highways to that of city environment. The node prediction design and routing algorithm also therefore need to adapt for these changes. Highway mobility model, which is essentially a one-dimensional model, is rather simple and easy to predict. But for city mobility model, street structure, variable node density, presence of buildings and trees that behave as obstacles to even small distance communication make the model application that very complex and difficult.

Delay Constraints

The safety aspect (such as accidents, brake event) of VANET application warrants on time delivery of message to relevant nodes. It simply cannot compromise with any hard data delay in this regard. Therefore high data rates are not as important an issue for VANET as overcoming the issues of hard delay constraints.

Interaction with onboard sensors

This sensors helps in providing node location and their movement nature that are used for effective communication link and routing purposes.

Battery power and storage capacity

In modern vehicles battery power and storage is unlimited. Thus it has enough computing power which is unavailable in MANET. It is helpful for effective communication & making routing decisions.

Applications

The VANET application can be divided into two major categories [4]:

Safety

Safety applications have the ability to reduce traffic accidents and to improve general safety. These can be further categorized as safety-critical and safety-related applications. In the design of security, it should be made sure safety messages are not forged.

Safety-critical

These are used in the case of hazardous situations (e.g. like collisions) [5]. It includes the situations where the danger is high or danger is imminent [6]. Safety-critical applications involve communication between vehicles (V2V) or between vehicles and infrastructure/infrastructure and vehicles (V2I/I2V).

Safety-related

These include safety applications where the danger is either low (curve speed warning) or elevated (work zone warning), but still foreseeable [6]. In safety-related applications, the latency requirements are not as stringent as in the case of safety-critical ones. Safety-related applications can be V2V or V2I/I2V.

Non-safety

These are applications that provide traffic information and enhance driving comfort. Non-safety applications mostly involve a V2I or I2V communication [4][5]. These services access the channels in the communication system, except the control channel. They access the channel in a low priority mode compared to safety applications.

Traffic optimization

Traffic information and recommendations, enhanced route guidance etc.

Infotainment

The Infotainment services are Internet access, media downloading, instant messaging etc.

Payment services

Payment services like Electronic toll collection, parking management etc.

Roadside service finder

Finding nearest fuel station, restaurants etc. This involves communication of vehicles with road side infrastructure and the associated database.

VANET Routing Protocols

In VANET, the routing protocols are classified into five categories: Topology based routing protocol, Position based routing protocol, Cluster based routing protocol, Geo cast routing protocol and Broadcast routing protocol. These protocols are characterized on the basis of area / application where they are most suitable. Fig. 1 shows the different routing protocols in VANET.

Topology Based Routing Protocols

These routing protocols use links information that exists in the network to perform packet forwarding. They are further divided into Proactive and Reactive.

Proactive routing protocols

The proactive routing means that the routing information, like next forwarding hop is maintained in the background irrespective of communication requests. The advantage of proactive routing protocol is that there is no route discovery since the destination route is stored in the background, but the disadvantage of this protocol is that it provides low latency for real time application. A table is constructed and maintained within a node. So that, each entry in the table indicates the next hop node towards a certain destination. It also leads to the maintenance of unused data paths, which causes the reduction in the available bandwidth. The various types of proactive routing protocols are: LSR, FSR.

Reactive/Ad hoc based routing

Reactive routing opens the route only when it is necessary for a node to communicate with each other. It maintains only the routes that are currently in use, as a result it reduces the burden in the network. Reactive routing consists of route discovery phase in which the query packets are flooded into the network for the path search and this phase completes when route is found. The various types of reactive routing protocols are AODV, PGB, DSR and TORA

Position Based Routing Protocols

Position based routing consists of class of routing algorithm. They share the property of using geographic positioning information in order to select the next forwarding hops. The packet is send without any map knowledge to the one hop neighbour which is closest to destination. Position based routing is beneficial since no global route from source node to destination node need to be created and maintained. Position based routing is broadly divided in two types: Position based greedy V2V protocols, Delay Tolerant Protocols.

Cluster Based Routing

Cluster based routing is preferred in clusters. A group of nodes identifies themselves to be a part of cluster and a node is designated as cluster head will broadcast the packet to cluster. Good scalability can be provided for large networks but network delays and overhead are incurred when forming clusters in highly mobile VANET. In cluster based routing virtual network infrastructure must be created through the clustering of nodes in order to provide scalability. The various Clusters based routing protocols are COIN and LORA_CBF.

Geo Cast Routing

Geo cast routing is basically a location based multicast routing. Its objective is to deliver the packet from source node to all other nodes within a specified

geographical region (Zone of Relevance ZOR). In Geo cast routing vehicles outside the ZOR are not alerted to avoid unnecessary hasty reaction. Geo cast is considered as a multicast service within a specific geographic region. It normally defines a forwarding zone where it directs the flooding of packets in order to reduce message overhead and network congestion caused by simply flooding packets everywhere. In the destination zone, unicast routing can be used to forward the packet. One pitfall of Geo cast is network partitioning and also unfavorable neighbors, which may hinder the proper forwarding of messages. The various Geo cast routing protocols are IVG, DG-CASTOR and DRG.

Broadcast Routing

Broadcast routing is frequently used in VANET for sharing, traffic, weather and emergency, road conditions among vehicles and delivering advertisements and announcements. The various Broadcast routing protocols are BROADCOMM, UMB, V-TRADE, and DV-CAST.

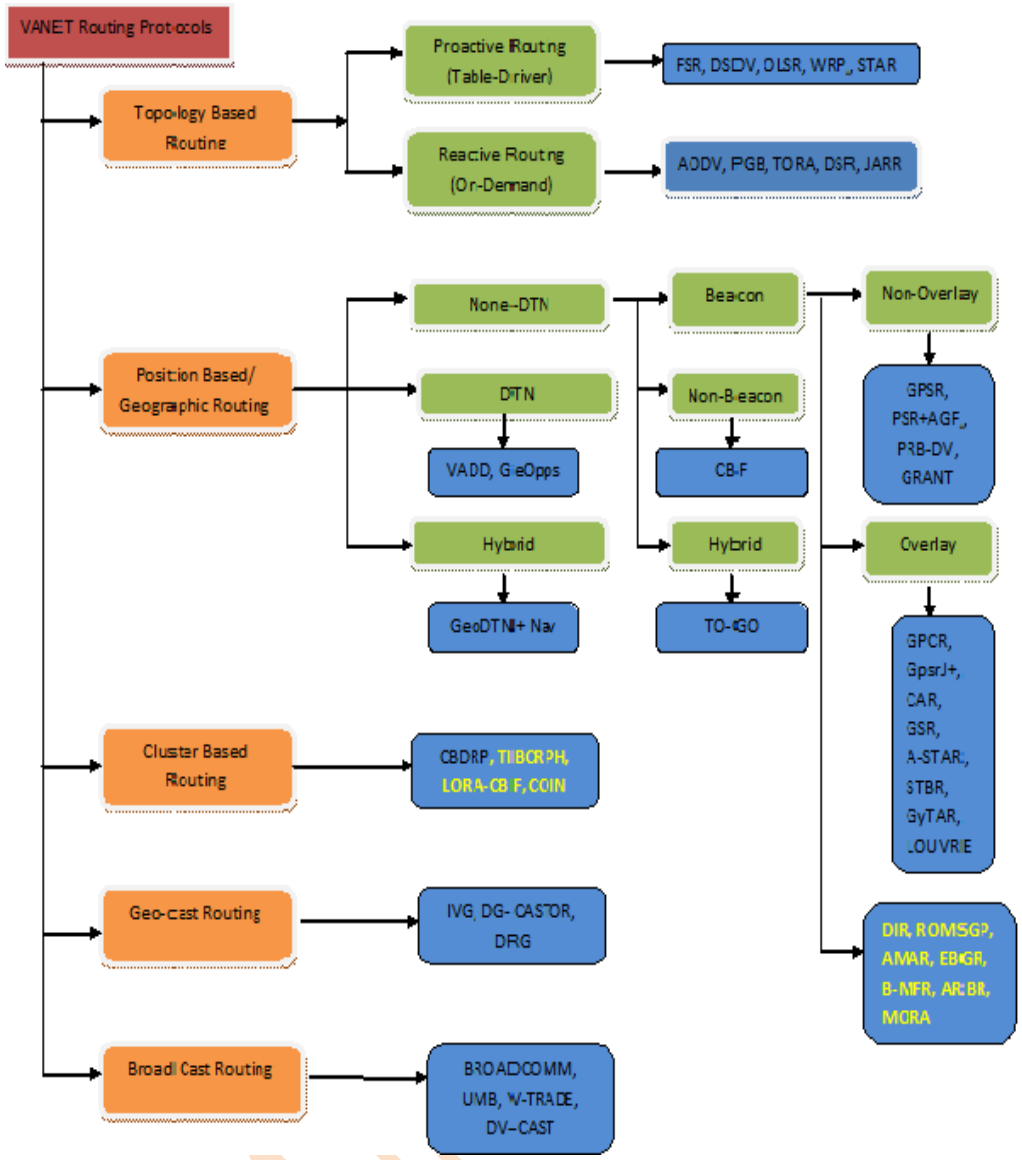


Fig. 1 Routing Protocols in VANET

Selected Proactive and reactive protocols

In this paper, an attempt has been made to compare three well know protocols AODV, AOMDV, DSR by using three performance metrics packet delivery ratio and average end to end delay.

Ad hoc On-Demand Distance Vector (AODV) Routing Protocol

In AODV[7] (Perkins, 1999) routing, upon receipt of a broadcast query (RREQ), nodes record the address of the node sending the query in their routing table. This procedure of recording its previous hop is called *backward learning*. Upon arriving at the destination, a reply packet (RREP) is then sent through the complete path obtained from backward learning to the source. The AODV algorithm enables dynamic, self-starting, multihop routing between

participating mobile nodes wishing to establish and maintain an ad hoc network. AODV allows mobile nodes to obtain routes quickly for new destinations, and does not require nodes to maintain routes to destinations that are not in active communication. AODV allows mobile nodes to respond to link breakages and changes in network topology in a timely manner. The operation of AODV is loop-free, and by avoiding the Bellman-Ford "counting to infinity" problem offers quick convergence when the adhoc network topology changes (typically, when a node moves in the network). When links break, AODV causes the affected set of nodes to be notified so that they are able to invalidate the routes using the lost link. Route Requests (RREQs), Route Replies (RREPs) and Route Errors (RERRs) are message types defined by AODV [7].

Ad hoc On-demand Multipath Distance Vector (AOMDV)

AOMDV[8] protocol is an extension based on Ad hoc On demand Distance Vector (AODV). However, the performance of AOMDV is much better than AODV [2]. AOMDV can find node-disjoint paths and link-disjoint paths when discovering routes. Because the conditions of node-disjoint paths are much stricter than that of link-disjoint paths, the number of node-disjoint paths is less than that of link-disjoint paths. Thus link-disjoint policy is used more popular. After multiple paths are found, AOMDV will store the paths in routing table. The source node will select one established path according to the timestamp. The first selected forward path is the earliest established one. For route maintenance, when a route failure is detected, packets can be forwarded through other paths. To ensure the freshness of routes, timeout mechanism is adopted. The HELLO messages are broadcasted to eliminate expired routes.

As well as AODV, AOMDV is an on-demand routing protocol. When a source node needs a route to a destination, and there are not available paths, the source node will broadcast RREQ routing packet to initiate a route discovery process. Other nodes may receive duplicate RREQ packets due to flooding. When this case occurs, other nodes will establish or update multiple reverse paths according to different first hops of RREQ packets. However, AODV will establish a reverse path using the first RREQ packet and other duplicate RREQ packets are discarded. After reverse paths establishing, intermediate nodes will search their routing tables for an available forward path to destination node. If the path exists, an RREP packet will be sent back to source node along a reverse path and the RREQ packet will be discarded. If the path does not exist and the intermediate node does not forward other duplicate RREQ packets, the RREQ packet will be broadcasted. When destination node receives RREQ packet, it will establish or update reverse paths, too. However, destination node will reply with looser policy to find multiple link disjoint paths. According to the reply policy, the destination

node will reply all RREQ packets from different neighbors although the RREQ packets posses same first hop. Different RREP packets will be sent back through different neighbors, which can ensure link-disjoint path establishment. After passing by different neighbors, RREQ packets will be sent to source node along link-disjoint reverse paths. When intermediate and source nodes receive RREP packets, they will establish loop-free and link-disjoint paths to destination node according to different first hops of RREP packets. For intermediate nodes that are shared by different link-disjoint paths, they will check if there are unused reverse paths to the source node. If so, one reverse path will be selected to forward the current RREP packet; otherwise, the packet will be discarded.

Dynamic Source Routing

The Dynamic Source Routing protocol (DSR) [9] is (Perkins, 2007), an on demand routing protocol. DSR is a simple and efficient routing protocol designed specifically for use in multi-hop wireless ad hoc networks of mobile nodes. Using DSR, the network is completely self-organizing and self-configuring, requiring no existing network infrastructure or administration. The DSR protocol is composed of two main mechanisms that work together to allow the discovery and maintenance of source routes in the ad hoc network:

Route Discovery is the mechanism by which a node S wishing to send a packet to a destination node D obtains a source route to D. Route Discovery is used only when S attempts to send a packet to D and does not already know a route to D.

Route Maintenance is the mechanism by which node S is able to detect, while using a source route to D, if the network topology has changed such that it can no longer use its route to D because a link along the route no longer works. When Route Maintenance indicates a source route is broken, S can attempt to use any other route it happens to know to D, or it can invoke Route Discovery again to find a new route for subsequent packets to D. Route Maintenance for this route is used only when S is actually sending packets to D.

In DSR Route Discovery and Route Maintenance each operate entirely" on demand".

Simulation Based Analysis using Network Simulator (NS-2)

In this section we have described about the tools and methodology used in our paper for analysis of adhoc routing protocol performance i.e. about simulation tool, Simulation Setup(traffic scenario, Mobility model) performance metrics

used and finally the performance of protocols is represented by using excel graph.

Simulation Tool

In this paper the simulation tool used for analysis is NS-2[10] which is highly preferred by research communities. NS is a discrete event simulator targeted at networking research. Ns provides substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks. NS2 is an object oriented simulator, written in C++, with an OTcl interpreter as a frontend. This means that most of the simulation scripts are created in Tcl(Tool Command Language). If the components have to be developed for ns2, then both tcl and C++ have to be used.

Simulation Setup

The table1 below list the details of simulation setup used in this simulation based analysis.

Platform	Windows Vista Ultimate (using Cygwin 1.7)
NS version	Ns –allinone-2.29
Simulation time	300 s
Topology size	4000 m x 7000 m
Routing Protocols	AODV, AOMDV, DSR
Traffic Type	TCP
Data type	CBR
Data Packet Size	512 bytes
MAC protocol	IEEE 802.11
Radio Propagation Model	Two Ray Ground

Table 1: Simulation Setup

Simulation Metrics used

The following metrics are used in this paper for the analysis of AODV, AOMDV and DSR routing protocols.

Packet Delivery Ratio (PDR): ***It is the fraction of generated packets by received packets. That is, the ratios of packets received at the destination to those of the packets generated by the source. As of relative amount, the usual calculation of this system of measurement is in percentage (%) form. Higher the percentage, more privileged is the routing protocol.***

Average End-to-End Delay (E2E Delay):

It is the calculation of typical time taken by packet (in average packets) to cover its journey from the source end to the destination end. In other words, it covers all of the potential delays such as route discovery, buffering processes, various in-between queuing stays, etc, during the entire trip of transmission of the packet. The classical unit of this metric is millisecond (ms). For this metric, lower the time taken, more privileged the routing protocol is considered.

Simulation Results

Variable	Value
No. of nodes	12
Maximum Connections	8

Table 2 Connection pattern

Figure 2 represents the performance of AODV, AOMDV and DSR.

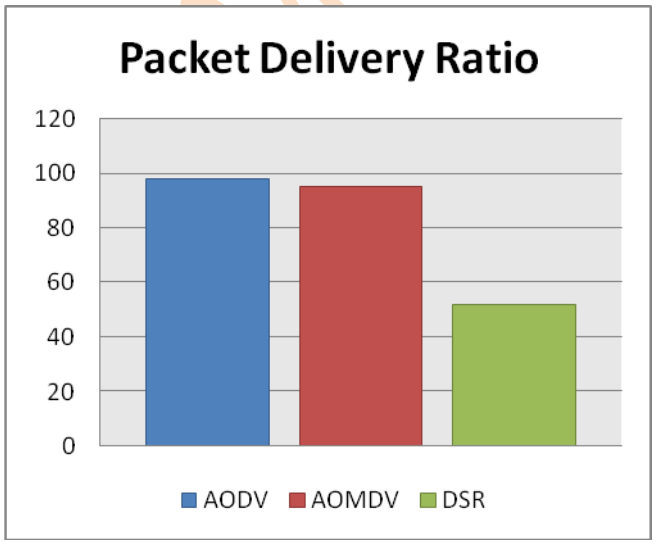


Fig. 2 PDR vs. Node Density at city low density

Figure 3 represents the performance of AODV, AOMDV and DSR in terms of Average End to End Delay vs. Node Low Density

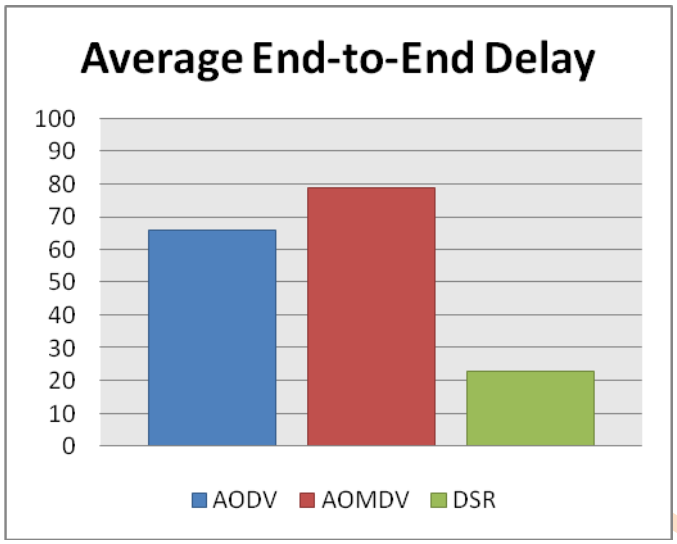


Fig. 3 Average E2E Delay (in ms) vs. Node Density at city low density

Conclusion

In this paper the analysis of adhoc routing protocol is done in realistic scenario of VANET. After doing the simulation based analysis of AODV, AOMDV and DSR in realistic scenario of VANET we can see that the performance of AODV in terms of PDR is very good approximate 98%. The Average end to end delay of AODV is very high. The DSR performs well in both of the scenario in terms of Avg. end to end delay. Packet delivery Ratio of AODV is better than other two protocols so we can say this protocol is applicable to carry sensitive information in VANET but it fails for the scenario where transmission time should be very less as it has highest end to end delay. For quick transmission DSR performs well but not suitable to carry information as packet loss is very high. The performance of AOMDV is average.

References

[1] Y. Zang, L. Stibor, and H. J. Reumerman, “Neighborhood evaluation of vehicular ad-hoc network using IEEE 802.11p,” in Proceedings of the 8th European Wireless Conference, p. 5, Paris, France, 2007.

[2] IEEE Draft P802.11p/D2.0, November 2006. Wireless Access in Vehicular Environments (WAVE).

[3] Hartenstein, H., et al., "Position-Aware Ad Hoc Wireless Networks for Inter-Vehicle Communications: The FleetNet Project," *In Proceedings MobiHoc'01: 2nd ACM Int'l. Symp. Mobile Ad Hoc Networking & Computing*, New York: ACM Press, pp. 259–62, 2001.

[4] Vehicle Safety Communications Project Task 3 Final Report. Technical report, The CAMP Vehicle Safety Communications Consortium, Mar 2005. Sponsored by U.S. Department of Transportation (USDOT). Available through National Technical Information Service, Springfield, Virginia 22161.

[5] Rainer Kroh, Antonio Kung, and Frank Kargl. VANETS Security Requirements Final Version. Technical report, Secure Vehicle Communication (Sevecom), Sep 2006. Available at http://www.sevecom.org/Pages/Project Documents.html.

[6] Elmar Schoch, Frank Kargl, Michael Weber, and Tim Leinmuller. Communication Patterns in VANETs. IEEE Communications Magazine, 46:119–125, Nov 2008.

[7] C.E. Perkins and E.M. Royer. Ad-hoc on-demand distance vector routing. In *Proceedings*of the 2nd IEEE Workshop on Mobile Computing Systems and Applications, pages 90–100, 1999.

[8] Y. Yuan, H. Chen, and M. Jia, "An optimized ad-hoc on-demand multipath distance vector (AOMDV) routing protocol," Proc. 11th Asia-Pacific Conference on Communications, IEEE Press, Oct. 2005, pp. 569-573, doi:10.1109/APCC.2005.1554125.

[9] David B. Johnson and David A. Maltz. Dynamic source routing in ad hoc wireless networks.In *Mobile Computing*, volume 353. 1996.

[10] Sandeep Gupta. "A brief guide to ns2."

IT 003

AN APPROACH ON PREPROCESSING OF DATA STREAMS

Mr.Avinash L Golande
avinash.golande@gmail.com
Ph.9890915321

Abstract - *T*he recent advances in hardware and software have enabled the capture of different measurements of data in a wide range of fields. These measurements are generated continuously and in a very high fluctuating data rates. Examples include sensor networks, web logs, and computer network traffic. The storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non stopping streams of information. The research in data stream mining has gained a high attraction due to the importance of its applications and the increasing generation of streaming information. Applications of data stream analysis can vary from critical scientific and astronomical applications to important business and financial ones. Algorithms, systems and frameworks that address streaming challenges have been developed over the past three years. In this review paper, we present the state of- the-art in this growing vital field.

Keywords-component, data stream, *VFDT*

Introduction

The intelligent data analysis has passed through a number of stages. Each stage addresses novel research issues that have arisen. Statistical exploratory data analysis represents the first stage. The goal was to explore the available data in order to test a specific hypothesis. With the advances in computing power, machine learning field has arisen. The objective is to find computationally efficient solutions to data analysis problems. Along with the progress in machine learning research, new data analysis problems have been addressed. Due to the increase in database sizes, new algorithms have been proposed to deal with the scalability issue. Moreover machine learning and statistical analysis techniques have been adopted and modified in order to address the problem of very large databases. Data mining is that interdisciplinary field of study that can extract models and patterns from large amounts of information stored in data repositories.

Recently, the data generation rates in some data sources become faster than ever before. This rapid generation of continuous streams of information has challenged our storage, computation and communication capabilities in

computing systems. Systems, models and techniques have been proposed and developed over the past few years to address these challenges.

In this paper, we review the theoretical foundations of data stream analysis, mining data stream systems, techniques are critically reviewed. Finally, we outline and discuss research problems in streaming mining field of study. These research issues should be addressed in order to realize robust systems that are capable of fulfilling the needs of data stream mining applications. The paper is organized as follows. Section 2 presents the theoretical background of data stream analysis. In sections 3 and 4 mining data stream techniques and systems are reviewed respectively. Open and addressed research issues in this growing field are discussed in section 5. Finally section 6 summarizes this review paper. section 7 enlist the references.

Theoretical Foundations

Research problems and challenges that have been arisen in mining data streams have its solutions using well established statistical and computational approaches. We can categorize these solutions to data-based and task-based ones. In data-based solutions, the idea is to examine only a subset of the whole dataset or to transform the data vertically or horizontally to an approximate smaller size data representation. At the other hand, in task-based solutions, techniques from computational theory have been adopted to achieve time and space efficient solutions. In this section we review these theoretical foundations

Data-based Techniques

Data-based techniques refer to summarizing the whole dataset or choosing a subset of the incoming stream to be analyzed. Sampling, load shedding and sketching techniques represent the former one. Synopsis data structures and aggregation represent the later one. Here is an outline of the basics of these techniques with pointers to its applications in the context of data stream analysis.

Sampling : Sampling refers to the process of probabilistic choice of a data item to be processed or not.
Load Shedding ItLoad shedding refers to the process of dropping a sequence of data streams.
Sketching:Sketching is the process of randomly project a subset of the features.
Synopsis Data Structures: Creating synopsis of data refers to the process of applying summarization techniques that are capable of summarizing the incoming stream for further analysis.

Aggregation:Aggregation is the process of computing statistical measures such as means and variance that summarize the incoming stream.

Task-based Techniques

Task-based techniques are those methods that modify existing techniques or invent new ones in order to address the computational challenges of data stream processing. Approximation algorithms, sliding window and algorithm output granularity represent this category. In the following subsections, we examine each of these techniques and its application in the context of data stream analysis.

Approximation algorithms – they have their roots in algorithm design. It is concerned with design algorithms for computationally hard problems.

Sliding Window - The inspiration behind sliding window is that the user is more concerned with the analysis of most recent data streams.

Algorithm Output Granularity -The algorithm output granularity (AOG) introduces the first resource-aware data analysis approach that can cope with fluctuating very high data rates according to the available memory and the processing speed represented in time constraints.

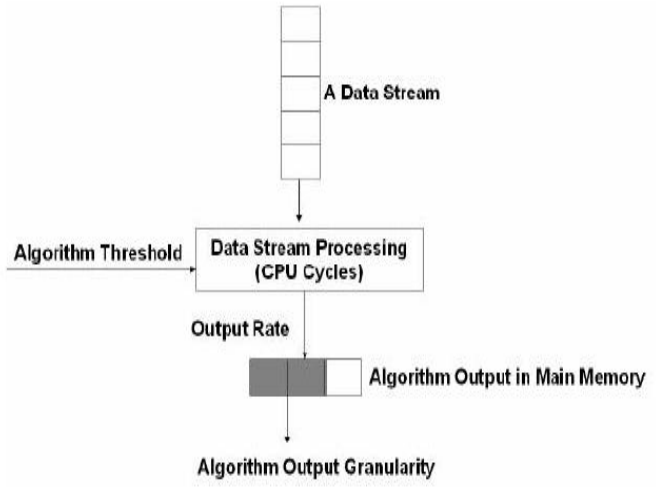


Fig1 The AOG Algorithm

Mining Techniques

Mining data streams has attracted the attention of data mining community for the last three years. Number of algorithms has been proposed for extracting knowledge from streaming information. In this section, we review clustering, classification, frequency counting and time series analysis techniques.

Clustering : Guha et al. have studied analytically clustering data streams using K-median technique. Charikar et al have proposed another K- median algorithm that overcomes the problem of increasing approximation factors in the Guha et al algorithm with the increase in the number of levels used to result in the final solution of the divide and conquer algorithm. The algorithm has also been studied analytically. Domingos et al. have proposed a general method for scaling up machine learning algorithms. They have termed this approach Very Fast Machine Learning VFML.

Classification : Wang et al. have proposed a general framework for mining concept drifting data streams. Ganti et al. have developed analytically an algorithm for model maintenance under insertion and deletion of blocks of data records. This algorithm can be applied to any incremental data mining model.

GEMM algorithm accepts a class of models and an incremental model maintenance algorithm for the unrestricted window option, and outputs a model maintenance algorithm for both window-independent and window dependent block selection sequence. Domingos et al. have developed VFDT. It is a decision tree learning systems based on Hoeffding trees.

Frequency Counting :

Giannella et al. have developed a frequent itemsets mining algorithm over data stream. They have proposed the use of tilted windows to calculate the frequent patterns for the most recent transactions based on the fact that users are more interested in the most recent transactions.

Cormode and Muthukrishnan have developed an algorithm for counting frequent items. The algorithm uses group testing to find the hottest k items. Gaber et al. have developed one more AOG-based algorithm: Lightweight frequency counting LWF. It has the ability to find an approximate solution to the most frequent items in the incoming stream using adaptation and releasing the least frequent items regularly in order to count the more frequent ones.

Time Series Analysis

Indyk et al. have proposed approximate solutions with probabilistic error bounding to two problems in time series analysis: relaxed periods and average trends. Perlman and Java have proposed a two phase approach to mine astronomical time series streams.

Zhu and Shasha have proposed techniques to compute some statistical measures over time series data streams. Lin et al. have proposed the use of symbolic representation of time series data streams. This representation allows dimensionality/numerosity reduction. They have demonstrated the applicability of the proposed representation by applying it to clustering, classification, indexing and anomaly detection. The approach has two main stages. The first one is the transformation of time series data to Piecewise Aggregate Approximation followed by transforming the output to discrete string symbols in the second stage.

SYSTEMS

Several applications have stimulated the development of robust streaming analysis systems. The following represents a list of such applications. · Burl et al. have developed *Diamond Eye* for NASA and JPL. Kargupta et al. have developed the first ubiquitous data stream mining system: *Mobilize*. It is a client/server PDA-based distributed data stream mining application for stock market data. It should be pointed out that the mining component is located at the server side rather than the PDA. There are different interactions between the server and PDA till the results finally displayed on the PDA screen. The tendency to perform data mining at the server side has been changed with the increase of the computational power of small devices.

RESEARCH ISSUES

Data stream mining is a stimulating field of study that has raised challenges and research issues to be addressed by the database and data mining communities. The following is a discussion of both addressed and open research issues. The following is a brief discussion of previously addressed issues is as follows:

Unbounded memory requirements due to the continuous flow of data streams:

Required result accuracy:

Transferring data mining results over a wireless network with a limited bandwidth:

Modeling changes of mining results over time

Developing algorithms for mining results' changes

Visualization of data mining results on small screens of mobile devices

Interactive mining environment to satisfy user requirements

The integration between data stream management systems and the ubiquitous data stream

mining approaches:

The needs of real world applications:

Data stream pre-processing:

Model over fitting:

Data stream mining technology

The formalization of real-time accuracy evaluation

Research Issues	**Challenges**	**Approaches**
Memory Management	Fluctuated and irregular data arrival rate and variant data arrival rate over time	Summarizing techniques
Data preprocessing	Quality of mining results and automation of preprocessin	Light-weight preprocessing techniques

	g techniques	
Compact data structure	Limited memory size and large volume of data streams	Incremental maintaining of data structure, novel indexing, storage and querying techniques
Resource aware	Limited resources like storage and computation capabilities	AOG
Visualization of results	Problems in data analysis and quick decision making by user	Still is a research issue (one of the proposed approaches is: intelligent monitoring)

Table 1 Classification of data stream mining challenges [8]

Conclusions

The dissemination of data stream phenomenon has necessitated the development of stream mining algorithms. The area has attracted the attention of data mining community. The proposed techniques have their roots in statistics and theoretical computer science. Data-based and task-based techniques are the two categories of data stream mining algorithms. Based on these two categories, a number of clustering, classification, and frequency counting and time series analysis have been developed. Systems have been implemented to use these techniques in real applications. Mining

data streams is still in its infancy state. Addressed along with open issues in data stream mining are discussed in this paper. Further developments would be realized over the next few years to address these problems. Having these systems that address the above research issues developed, that would accelerate the science discovery in physical and astronomical applications , in addition to business and financial ones that would improve the real-time decision making process.

REFERENCES

[1] Mining Data Streams: A Review Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy Centre for Distributed Systems and Software Engineering, Monash University

[2] C. Aggarwal, J. Han, J. Wang, P. S. Yu, A Framework for Clustering Evolving Data Streams, Proc. 2003 Int. Conf. on Very Large Data Bases, Berlin, Germany, Sept. 2003.

[3] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, On Demand Classification of Data Streams, Proc. 2004 Int. 24 SIGMOD Record, Vol. 34, No. 2, June 2005 Conf. on Knowledge Discovery and Data Mining, Seattle, WA, Aug. 2004.

[4] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, A Framework for Projected Clustering of High Dimensional Data Streams, Proc. 2004 Int. Conf. on Very Large Data Bases, Toronto, Canada, 2004.

[5] DATA STREAM MINING - A Practical Approach Albert Bifet and Richard Kirkby, August 2009 Hebah H. O. NasereddinDepartment of computer Information system Faculty of IT Amman Arab University for Graduate Studies,Amman – Jordan

[6] An analytical framework for data stream mining techniques based on challenges and requirements Mahnoosh Kholghi (Department of Electronic, Computer and IT, Islamic Azad University, Qazvin Branch, Qazvin, Iran and member of Young Researchers Club), Mohammadreza Keyvanpour (Department of Computer Engineering Alzahra University Tehran, Iran)

[7] R. Bhargava, H. Kargupta, and M. Powers, Energy Consumption in Data Analysis for On-board and Distributed Applications, Proceedings of the ICML'03 workshop on Machine Learning Technologies for Autonomous Space Applications, 2003.

[8] Gaber, M.M., Krishnaswamy, S., and Zaslavsky, A. (2006). *On-board Mining of Data Streams in Sensor Networks*, In *Advanced Methods of Knowledge Discovery from Complex Data*, S. Badhyopadhyay, et al., Editors., Springer. pp. 307-335.

WebSites and reference Books:

[9] http://researcher.ibm.com

[10] http://domino.watson.ibm.com

IT 004

TEXT SEGMENTATION FOR COMPRESSION FOR MULTI-LAYERED MRC DOCUMENT

Mrs.Prajakta P. Bastawade

Department of Information Technology

MIT College of Engineering, Pune

prajaktamagdum@gmail.com

Contact no-9850330669

Mrs. Bharati Dixit

Head of Information Technology Department

MIT College of Engineering ,Pune

dixit.bharati@gmail.com

Contact no-9822015353

Abstract *— MRC uses a multi-layer, multi-resolution representation of a compound document. Instead of using a single algorithm, it uses multiple compression algorithms including the one specifically developed for text and images. The 3 layer MRC model contains 2 color image layers (foreground (FG) and background (BG)) and one binary image layer (mask). It improves the resulting quality and compression ratio of the complex document as compared to lossy image compression algorithms. Therefore, the important main step is the segmentation algorithm used to compute the binary mask.*

Keywords — Document compression, Image segmentation MRC compression, Text Segmentation.

I. INTRODUCTION

With the large use of processing electronic imaging and scanning devices it is necessary to efficiently compress, store, and transfer large document files. JPEG and JPEG2000 are frequently used tool, are not very effective for the compression of raster scanned compound documents which contains combination of text, graphics, and natural images. The mixed raster content (MRC) specifies framework for layer-based document compression defined in the ITU-T T.44 [1] that enables the preservation of text detail while reducing

the bitrate of encoded raster documents. The most traditional approach to document binarization is Otsu's method [2] which thresholds pixels in an effort to divide the document's histogram into object and background. Many text segmentation approaches have been based upon statistical models. One of the best commercial text segmentation algorithms, which is incorporated in the DjVu document encoder, uses a hidden Markov model (HMM) [3].

This paper introduces a proposed multiscale segmentation algorithm for both detecting and segmenting text from compound document which consist of combination of text, images & graphics. The segmentation algorithm consists of two algorithms which are to be applied in sequence. Cost Optimized Segmentation (COS) algorithm and Connected Component Classification (CCC) algorithm. The third algorithm is incorporation of COS/CCC algorithms into a multiscale framework to improve detection of varying size text.

Section II describes basics of MRC standard, Section III and Section IV describes proposed COS and CCC algorithm, Section V describes proposed MULTISCALE-COS/CCC algorithm .Section VI describes experimental results.

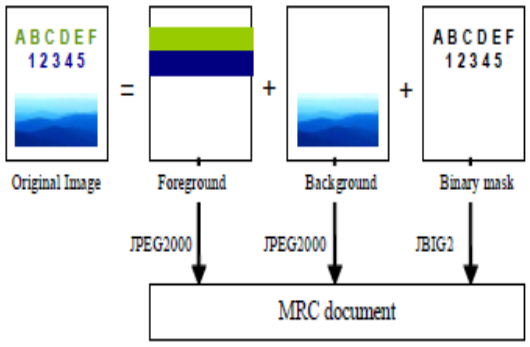
II. MRC DOCUMENT COMPRESSION STANDARD

In MRC original document is divided into three layers: a binary mask layer, foreground layer, and background layer. The binary mask indicates the assignment of each pixel to the foreground layer or the background layer by a "1" or "0" value, respectively. Typically, text regions are classified as foreground while picture regions are classified as background. Each layer is then encoded independently using an appropriate encoder. Foreground and background layers may be encoded using traditional photographic compression algorithm such as JPEG or JPEG2000 while the binary mask layer may be encoded using symbol-matching based compression such as JBIG or JBIG2. Typically, the foreground layer is more aggressively compressed than the background layer because the foreground layer requires lower color and spatial resolution. Fig. 1 shows an example of layers in an MRC mode 1 document [4]

The most important step in MRC is segmentation step which creates a binary mask that separates text and line-graphics from image and background regions in the document. Text

segmentation is extracting text components from a document Segmentation affects both the quality and bitrate of an MRC document.

Fig .1 Model of MRC Document



III. COST OPTIMIZE SEGMENTATION (COS)

COS is used for initial segmentation which is formulated in a global cost optimization framework. The COS algorithm is a blockwise segmentation algorithm based on cost optimization. The COS algorithm[4] creates a binary image from a gray level or color document The resulting binary image contains many false text detections (non text components) which is further processed by CCC algorithm to improve the accuracy of the segmentation.

Fig 2 . COS Algorithm

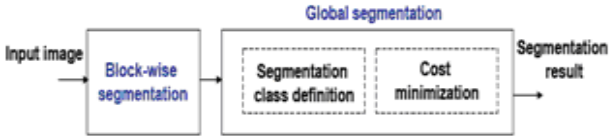


Fig.2 shows Cost Optimized Segmentation (COS) algorithm. It comprises two steps: Blockwise segmentation and Global segmentation.

A. BLOCK-WISE SEGMENTATION

Blockwise segmentation is performed by first dividing the image into overlapping blocks. Each block segmented independently using clustering procedure .Each block contains $m \times m$ pixels and adjacent blocks overlap by $m/2$ pixels in both horizontal and vertical directions. The blocks are denoted, $O_{i,j}$ for $i = 1, .., M$, and j = $1, .., N$, where M and N are the number of the blocks in the vertical and horizontal directions. The pixels in each block are segmented into foreground ("1") or background ("0") by the clustering method of Cheng and Bouman [5] This results in an initial binary mask for each lock denoted by $C_{i,j} \in \{0, 1\}$ m×m.

B. GLOBAL SEGMENTATION

However, in order to form a consistent segmentation of the page, these initial block segmentations must be merged into a single binary mask in global segmentation .After initial block segmentation, four possible classes are defined for each block .we allow each block to be modified using a class assignment, $S_{i,j} \in \{0, 1, 2, 3\}$, as follows

Class 0: Original segmentation

Class 1: Reversed

Class 2: All background

Class 3: All foreground

If the block class is "original," then the original binary segmentation of the block is retained. If the block class is "reversed," then the assignment of each pixel in the block is reversed (i.e., 1 goes to 0, or 0 goes to 1). If the block class is set to "all background" or "all foreground," then the pixels in the block are set to all 0's or all 1's, respectively.

Our objective is then to select the class assignments, $S_{i,j} \in \{0, 1, 2, 3\}$, so that the resulting binary masks, $\tilde{C}_{i,j}$, are consistent. We do this by minimizing the following global cost as a function of the class assignments

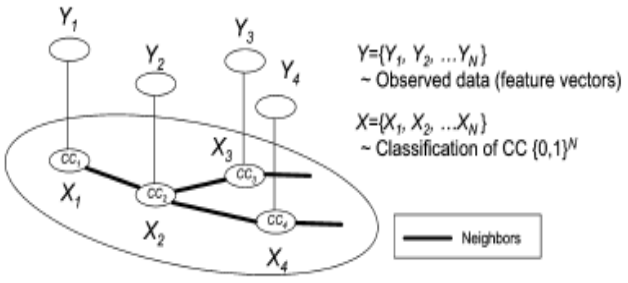
$$Cost(S) = \sum_{i=0}^{M} \sum_{i=0}^{N} \left\{ E(s_{i,j}) + \lambda_1 V_1(s_{i,j-1}, s_{i,j}) + \lambda_2 V_2(s_{i-1,j}, s_{i,j}) + \lambda_3 V_3(s_{i,j}) \right\} \quad (1)$$

$s_{i,j}$: Class of block at location (i,j). $S=\{s_{i,j}\}$
E : Total variance of gray levels of each group (0 or 1)
V_1 : Number of mismatches in horizontal overlap region
V_2 : Number of mismatches in vertical overlap region
V_3 : Number of '1' pixels inside block
λ_k : Weight coefficients, k=1,2,3

IV. CONNECTED COMPONENT CLASSIFICATION (CCC)

CCC[4] Refines COS results by removing false detections (non-text components) using Bayesian text detection procedure as shown in Fig 3.It operates on the binary image produced by COS. Line segments indicate dependency between random variables. Each component CC_i has an observed feature vector Y_i and a class label X_i. Neighboring pairs are indicated by thick line segments.

Fig 3. Bayesian segmentation model.



The CCC algorithm works in 3 steps as follows-

1. Extracting foreground connected components using a 4-point neighbourhood search.

2. Next a feature vector Yi, is calculated for the i^{th} connected component (CC_i). Each Y_i is a 4 dimensional feature vector which describes aspects of the i^{th} connected component such as edge depth and color uniformity.

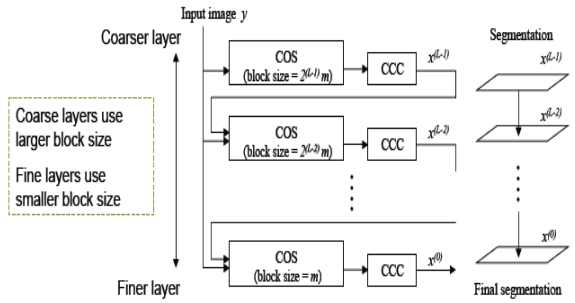
3. Each connected component also has a label xi, which is 1 if the component is text and 0 if it is non-text.

V. MULTISCALE-COS/CCC SEGMENTATION ALGORITHM

To improve accuracy for the detection of text with varying size, COS/CCC algorithms are incorporated into a multiscale framework [6] as shown in fig 3. It allows detecting both large and small components. The multiscale-COS/CCC divides the segmentation process into several scales. Each scale is numbered from 0 to L − 1, where 0 is the finest scale and L − 1 is the coarsest scale. Segmentation is performed from coarse to fine scales, where the coarser scales use larger block sizes, and the finer scales use smaller block sizes. The

segmentation on each layer uses results of the previous layer. Both COS and CCC are performed on each scale.

Fig. 4 Multiscale-COS/CCCalgorithm



New term in the COS cost function represents the number of pixel mismatches between

current and previous layers

$$Cost(S^{(n)}) = \sum_{i=0}^{M}\sum_{j=0}^{N}\left\{E(s_{i,j}^{(n)}) + \lambda_1 V_1(s_{i,j-1}^{(n)}, s_{i,j}^{(n)}) + \lambda_2 V_2(s_{i-1,j}^{(n)}, s_{i,j}^{(n)}) + \lambda_3 V_3(s_{i,j}^{(n)}) + \underline{\lambda_4 V_4(s_{i,j}^{(n)}, x_{i,j}^{(n+1)})}\right\} \quad (2)$$

New term

The new term V4 enforces consistency with coarser segmentation results.

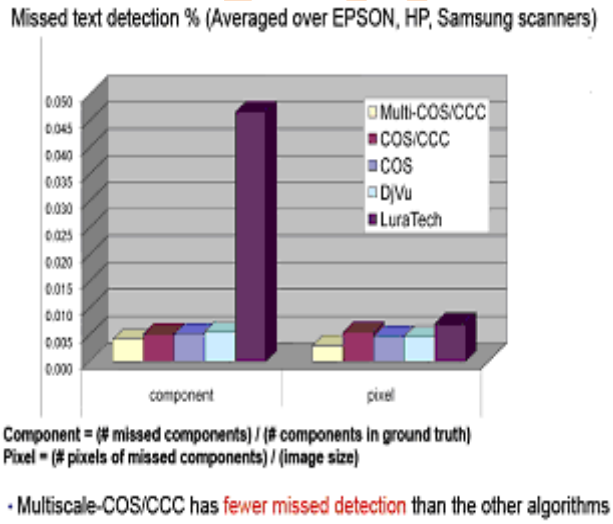
VI. Results

In this section, we compare multiscale-COS/CCC segmentation results with the results of two existing commercial software packages. Document Express ersion 5.1 with DjVu1 and LuraDocument PDF Compressor Desktop 2. Our comparison is primarily based on two aspects: the segmentation accuracy, and the bitrate resulting from JBIG2 compression of the binary segmentation mask.

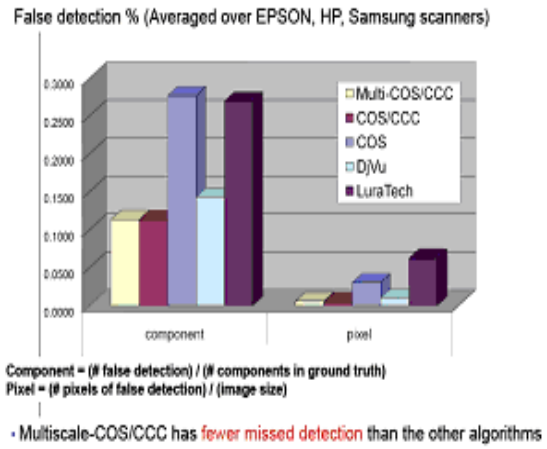
First, 38 documents were chosen from different document materials, including flyers, newspapers, and magazines. The 17 documents scanned by EPSON STYLUS PHOTO RX700 at 300 dpi were used for the training, and 21 documents scanned by EPSON STYLUS PHOTO RX700, HP photo smart 3300 All-in-One, and Samsung V SCX-5530FN at 300 dpi were used to verify the segmentation quality[4]

To evaluate the segmentation accuracy, the percent of missed detections and false detections of segmented components denoted as P_{MC} and P_{FC} are measured.

A) PERCENTAGE OF MISSED DETECTION



B) PERCENTAGE OF FALSE DETECTION



VII. CONCLUSION

Three algorithms for text segmentation: COS, COS/CCC, Multiscale-COS/CCC applied in sequence , accurately extract the text components as compared to commercial products .Also Robust over various paper materials, different scanner , and various image backgrounds and useful for applications such as Optical Character Recognition (OCR)

REFERENCES

[1] *ITU-T Recommendation T.44 Mixed Raster Content (MRC)*, T.44, International Telecommunication Union, 1999.

[2] Nobuyuki Otsu, "*A threshold selection method from gray- Level histograms*," IEEE trans. on Systems, Man and Cybernetics,vol. 9, no. 1, pp. 62–66, 1979.

[3] P. Haffner, L. Bottou, and Y. Lecun, "*A general segmentation scheme for DjVu document compression*," in *Proc. ISMM*, Sydney, Australia, Apr. 2002, pp. 17–36.

[4] Eri Haneda, and Charles A. Bouman, "*Text Segmentation for MRC Document Compression* ",IEEE Tansactions on Image Processing, Vol. 20, No. 6, June 2011

[5] H. Cheng and C. A. Bouman, "*Document compression using rate- Distortion optimized segmentation*," *J. Electron. Image.*, vol. 10, no.2, pp.460–474, 2001

[6] C. A. Bouman and B. Liu, "*Multiple Resolution segmentation of Textured Images*," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 2, pp. 99–113, 1991.

IT 005

SHOT BOUNDARY BASED KEY FRAME EXTRACTION TECHNIQUES: A SURVEY

M.A.Thalor[#1], Dr.S.T.Patil[#2]

[1,2]*Information Technology, University of Pune*
Pune. India
[1]*thalor.meenakshi@gmail.com*
[2]*stpatil77@gmail.com*

Abstract *- The key-frame extraction is the foundation of video analysis and content-based retrieval. Video key frames are pictures that can represent main content of the shot. In order to manage a video in a systematic manner, key frame extraction is essentially first step for facilitating the subsequent processes. Due to the importance of key frame, a fast and precise shot change detection method is always needed for many video-related applications including video browsing, video skimming, video indexing and video retrieval. Many algorithms have been proposed for detecting video shot boundaries .This paper presents a survey of several shot boundary detection techniques and their variations including pixel differences, histograms, discrete cosine transform and motion vector methods.*

Keywords*- **Key Feature Extraction and Shot Boundary Detection.***

1. INTRODUCTION

It is not efficient to process a video as whole so there is requirement of video segmentation. In video segmentation technique individual videos is converted into sequence of shots. From these shots extraction of key frames will carried out for subsequent processes. Figure 5.1 shows the hierarchical representation of video .The explanation of each term is as given below:

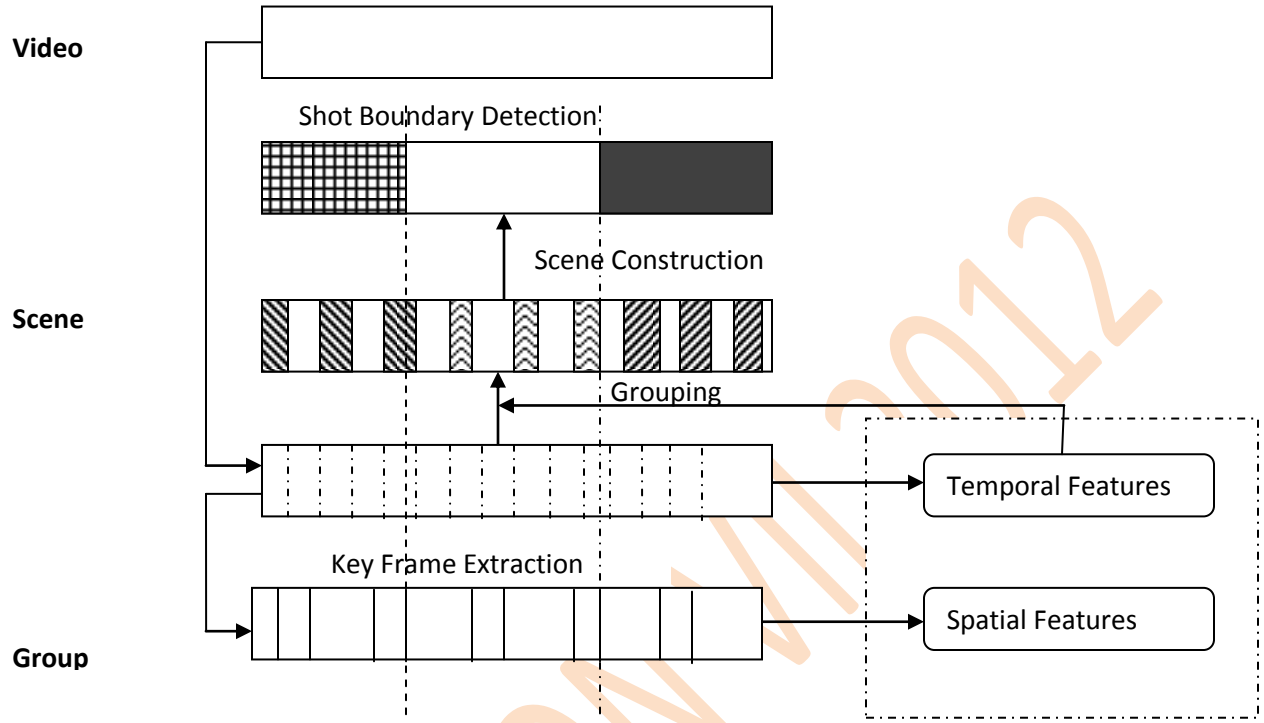


Figure 5.1: A Hierarchical Video Representation

A shot is defined as the consecutive frames from the start to the end of recording in a camera. It shows a continuous action in an image sequence. Key frame is the frame which represents the salient visual contents of a shot. Depending on the complexity of the content of the shot, one or more frame can be extracted. Video Scene is defines as collection of semantically related and temporally adjacent shots, depicting and conveying a high level concept or story. While shots are marked by physical boundaries, scenes are marked by semantic boundaries.

There are a number of different types of transitions or boundaries between shots. A cut is an abrupt shot change that occurs in a single frame. A fade is a slow change in brightness usually resulting in or starting with a solid black frame. A dissolve occurs when the images of the first shot get dimmer and the images of the second shot get brighter, with frames within the transition showing one image superimposed on the other. A wipe occurs when pixels from the second shot replace those of the first shot in a regular pattern such as in a line from the

left edge of the frames. Of course, many other types of gradual transition are possible.

2. RELATED WORK

In order to extract key frame from a video different approaches had been proposed some of these are listed below

1) Shot boundary based Approach

2) Visual content based approach

3) Motion analysis based approach

4) Shot activity based approach

5) Clustering based approach

Since a cut usually brings a huge color difference between the end of the current shot and the start of the next shot, a number of strategies have been suggested for the detection of shot changes. The major techniques that have been used for shot boundary detection are pixel differences, statistical differences, histogram comparisons, edge differences, compression differences, and motion vectors.

2.1 PIXEL DIFFERENCES

The cut between the two shots can be detected by counting the number of the intensity changed pixels [1]. The percentage of the total number of pixels changed exceeds a certain threshold Ti is calculated and compared to another threshold Tp. If this percentage is above Tp, a cut boundary is declared. However, noise, camera movement and moving objects can cause a large number of pixel changes and hence a false boundary will be detected. So this method is sensitive to camera and object motion.

2.2 STATISTICAL DIFFERENCES

Statistical methods expand on the idea of pixel differences by subdividing a video frame into blocks and then the blocks can be compared based on statistical characteristics of their intensity. This approach is better than the previous approach since it enhances the tolerance against the noise caused by camera or object movement but is slow due the complexity of the statistical formulas. However, it is also possible that the two corresponding blocks are different even though they have the same density function[2]. In other words, it also generates many false positives.

2.3 HISTOGRAMS

Histograms are the most common method used to detect shot boundaries. The simplest histogram method computes gray level or color histograms of the two images. The histograms of two frames will be almost the same if there is no significant camera motion involved. A shot boundary is detected if the bin-wise difference between the two histograms is above a threshold. Zhang, Kankanhalli, and Smoliar [1] compared pixel differences, statistical differences, and several different histogram methods and found that the histogram methods were a good trade-off between accuracy and speed. In order to properly detect gradual transitions such as wipes and dissolves, they used two thresholds. If the histogram difference fell between the thresholds, they tentatively marked it as the beginning of a gradual transition sequence, and succeeding frames were compared against the first frame in the sequence. If the running difference exceeded the larger threshold, the sequence was marked as a gradual transition.

2.4 COMPRESSION DIFFERENCES

As nowadays, most videos are stored in compressed format (e.g. MPEG), it is highly desirable for researchers to develop methods that can directly operate on the compressed stream. Arman et al. [2] shot boundaries by comparing a small number of connected regions. They used differences in the discrete cosine transform coefficients of JPEG compressed frames as their measure of frame similarity, thus avoiding the need to decompress the frames. Working on the compressed domain offers the many advantages. We don't have to perform fully decoding process thus the computational complexity and the decompression time can be reduced.

2.5 FEATURE BASED

Zabih[5] have detected the shot boundaries by looking for large edge change percentages. The percentage of edges that enter and exit between the two frames was computed. They determined that their method was more accurate at detecting cuts than histograms and much less sensitive to motion than chromatic scaling. In addition Zhang et. al. [1] proposes to use multiple visual criteria to extract key frames based on color and texture.

2.6 MOTION VECTORS

Ueda et. al.[3] and Zhang et.al.[1] used motion vectors determined from block matching to detect whether or not a shot was a zoom or a pan. Shahraray[4] used the motion vectors extracted as part of the region-based pixel difference computation described above to decide if there is a large amount of camera or object motion in a shot. Because shots with camera motion can be incorrectly classified as gradual transitions, detecting zooms and pans increases the accuracy of a shot boundary detection algorithm. Motion vector information can

also be obtained from MPEG compressed video sequences. However, the block matching performed as part of MPEG encoding selects vectors based on compression efficiency and thus often selects inappropriate vectors for image processing purposes.

.

3. CONCLUSION

Recall and precision are appropriate evaluation criteria which commonly used in the field of information retrieval. Recall is defined as the percentage of desired items that are retrieved. Precision is defined as the percentage of retrieved items that are desired items. The histogram algorithm gives consistent results. It usually produced the first or second best precision for a given recall value. The simplicity of the algorithm and the straight forward threshold selection make this algorithm a reasonable choice for many applications. Based on survey, pixel difference and statistical differences methods are comparatively slow than others and compression differences and motion vector gives expected result.

4. REFERENCES

[1] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "*Automatic partitioning of full-motion video,*" Multimedia Systems, pp.10–28, 1993.

[2] F. Arman, A. Hsu, and M-Y. Chiu, "*Image processing on encoded video sequences*" Multimedia Systems , pp. 211–219 ,1994.

[3] H. Ueda, T. Miyatake, and S. Yoshizawa, "*IMPACT: an interactive natural-motion-picture dedicated multimedia authoring system,*" Proc. CHI. 1991, pp. 343–350 ACM, New York ,1991.

[4] B. Shahraray, "*Scene change detection and content-based sampling of video sequences,*" in Digital Video Compression: Algorithms and Technologies, *Proc. SPIE* , pp.2–13 ,1995.

[5] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks," Proc. ACM Multimedia *95*, pp. 189–200, San Francisco, CA ,1995.

IT 006

Secure Mobile Ad hoc Network

Varsha S. Upare

Department of Computer Engg.

MIT College of Engg. Pune , varsha.upare@gmail.com

Abstract-

Mobile ad hoc network is the collection wireless mobile nodes that forms short term network. Security is an important in mobile ad hoc network.Manet is prone to security attack due to lack of centralized approach, dynamic network topology. In this paper, various attacks in the mobile ad hoc network have been studied. The malicious behaviour of a node and the need for the security in the ad hoc is defined in this paper. SCAN technology protects the network from data routing and data forwarding operrations.It guard the network from detecting and reacting to the malicious node. In SCAN each neighbouring node monitors each other, sustain each other, no node is better than other. The analysis of SCAN design has been studied in this paper.The survey analysis of packet delivery ratio,miss detection ratio with respect to mobilitiy has been covered.

KEYWORD-Mobile ad hoc network, Nodes, Security, Attack, Malicious Behaviour.

I. INTRODUCTION

A MANET is a self-organizing system of mobile nodes that communicate with each other via wireless links with no fixed infrastructure or centralized administration such as base stations or access points. MANETs are suitable for applications in which no infrastructure exists such as military battlefield, emergency rescue, vehicular communications and mining operations. In these applications, communication and collaboration among a given group of nodes are necessary. In this paper, study of the effects of different types of attacks on mesh-based multicast in MANETs is considered. Here, the most common types of attacks, namely rushing attack, black hole attack, neighbour attack and jellyfish attack is considered. Ad Hoc Network provides quick communication among nodes to transfer the packets from one node to other. All the links between nodes are wireless. Any malicious node in the network can disturb the whole process. Whenever a node exhibits a malicious behaviour under any attack, it assures the breach of security principles like availability, integrity, confidentiality etc [5]. An intruder takes advantage of the vulnerabilities ,presents in the ad hoc network and attacks the node which breaches the In this paper, we tackle an important security issue in ad hoc networks, namely the

protection of their network-layer operations from malicious attacks. We focus on securing the packet delivery functionality. Several recent studies [1]–[4] have provided detailed description on such network- layer security threats and their consequences. In SCAN, each node monitors the routing and packet forwarding behaviour of its neighbours, and independently detects any malicious nodes in its own neighbourhood. The monitoring mechanism takes advantage of the broadcast nature of wireless

Communication. In a network with reasonable node density, one node can often overhear the packets (including both routing updates and data packets)

Received, as well as the packets sent by a neighbouring node. In such cases, it can *cross-check* these packets to discover whether this neighbour behaves normally in advertising routing updates and forwarding data packets.

This paper is organized as follows: Section II covers the Various attacks in the mobile ad hoc network. Section III covers the vulnerabilities present in the ad hoc network. Due to this vulnerability, node behaves in malicious manner. Section IV defines malicious behaviour. . Section V describes the SCAN design in details. Section VI analyzes the miss detection ratio. False accusation and packet delivery ratio with respect to mobility. Section VII concludes the paper.

II. ATTACKS

Rushing attack.. When source nodes flood the network with route discovery packets in order to find routes to the destinations, each intermediate node processes only the first non-duplicate packet and discards any duplicate packets that arrive at a later time. Rushing attackers, by skipping some of the routing processes, can quickly forward these packets and be able to gain access to the forwarding group. This type of attacks was first introduced in [4].

• **Black hole attack**. In the balckhole attack, the attacker simply drops all of data packets it receives. This type of attack often results in very low packet delivery ratio. Simply forwards the packet without recording its ID in the packet, it makes two nodes that are not within th communication range of each other believe that they are neighbours, resulting in a disrupted route.

• **Jellyfish attack**. Similar to the black hole attack, a jellyfish attacker first needs to intrude into the forwarding group and then it delays data packets unnecessarily for some amount of time before forwarding them. This result in significantly high end-to-end delay and delay jitter, and thus degrades the performance of real-time applications. Jellyfish attacks were first discussed by Aad *et al.* in [1].

• **Neighbouring attack**. Upon receiving a packet, an intermediate node records its ID in the packet before forwarding the packet to the next node. However, if an

attacker simply forwards the packet without recording its ID in the packet, it makes two nodes that are not within the communication range of each other believe that they are neighbours resulting in a disrupted route.

III. NEED OF SECURITY IN AD HOC NETWORK

The security is important in mobile ad hoc network. Following are the some of the issues in the mobile ad hoc network.

Mobility- Each node in ad hoc network is movable. It can join or leave a network at any instant of time without informing any node. This gives chance to intruder to easily enter in the network or leave the network

Open Wireless Medium- All the communication among nodes is taking place through the medium of air an intruder can easily access medium to get

Information about the communication or can easily trap it.

Resource Constraint- Every node in mobile ad hoc network has limited resources like battery, computational power, bandwidth etc. An intruder can unnecessarily waste these limited resources in order to make it unavailable to perform.

Dynamic Network Topology- As the nodes are travel ,, the topology changes every time . The packets from source to destination may take different path for communication. An intruder can introduce itself in any path.

Scalability- Ad hoc network may contain of number of nodes. This number is not fixed. In a network of its range, as many as number of nodes can take part. Intruder simply takes advantage of this parameter as there is no limitation on number of nodes.

Reliability- All the wireless communication is limited to a range of 100 meter which puts a constraint on nodes to be in range for establishing communication. Due to this limited range, some data errors are also generated. For attacking a particular node, an intruder needs to be in its range.

IV. MALICIOUS BEHAVIOUR OF A NODE

Malicious Behaviour- "When a node breaks the security principles and is under any attack. Such nodes show one or more of the following behaviour:

Packet Drop- Simply consumes or drops the packet and does not forward it.

Battery Drained- A malicious node can waste the battery by performing unnecessarily operations.

Buffer Overflow- A node under attack can fill the buffer with forged updates so that real updates cannot be stored further.

Bandwidth Consumption- Whenever a malicious node consumes the bandwidth so that no other legitimate node can use it.

Malicious Node Entering- A malicious node can enter in the network without authentication.

Stale Packets- This means to inject stale packets into the network to create confusion in the network.

Link Break- malicious node restricts the two legitimate communicating nodes from communicating.

Message Tampering- A malicious node can alter the content of the packets.

Fake Routing- Whether there exists a path between nodes or not, a malicious node can send forged routes to the legitimate nodes in order to get the packets or to disturb the operations.

Stealing Information- Information like the content, location, sequence number can be stolen by the malicious node to use it further for attack.

Session Capturing- When two legitimate nodes communicate, a malicious node can capture their session so as to take some meaningful information.

V. SCAN DESIGN

In this section, the SCAN design has been studied in details. To secure the packet delivery functionality, each SCAN Node overhears the wireless channel in the promiscuous mode, and observes the routing and packet forwarding behaviour of its Neighbours at all time. A malicious node is convicted when its neighbours have reached such a consensus, then it is removed from the network membership and isolated in the network. To be a part of the network, each legitimate node carries a valid token. Without a valid token node cannot participate in the network. A legitimate node can always renew the token from its neighbours before its current Token expires. However, when a malicious node is convicted, its neighbours collectively revoke its current token and inform all other nodes in the network. Scan frame work goes through the following process.

- *Collaborative monitoring*: All nodes within a local neighbourhood collaboratively monitor each other.

- *Token renewal*: All legitimate nodes in a local neighbourhood collaboratively renew the tokens for each other.

- *Token revocation*: The neighbours of a malicious node upon consensus collaboratively revoke its current token. In this framework, the malicious nodes are detected and convicted via the collaborative monitoring mechanism.

A. Token Renewal

The token renewal mechanism Guarantees that legitimate nodes can persist in the Network by renewing their token from time to time. To participate in the network, each legitimate node carries a token which contains the following three fields (owner_time, signing_time, expiration_time). The tokens are protected by the public-key cryptographic mechanism. Before the current token expires, each node request its local neighbours to renew its token. The node that needs token renewal broadcasts a token request (TREQ) packet, which contains its current token and a timestamp. each node keeps a token revocation list (TRL) based on the token revocation mechanism. Specifically, when a node receives a TREQ packet from its neighbour, it removes the token from the packet. It checks whether the token has already been revoked by comparing it with the TRL. If the token is still valid yet about to expire, it constructs a new token with owner_id equal to that in the old token signing_time equal to the timestamp in the TREQ packet. the expiration_time is determined by the credit strategy defined as below.

1) Credit Strategy in Token Lifetime:

Here the credit strategy of a token is explained. In this strategy, a newly joined node is issued a token with short lifetime. It collect Its credit when it remains to behave well in the network and its subsequent token lifetime depends on its credit at the renewal time. The more credit one node has, the longer lifetime its token has. This way, a legitimate node will have its token lifetime steadily increased over time, thus renewing its token less and less frequently.

B. Collaborative Monitoring

The collaborative monitoring mechanism in SCAN observes the routing and packet forwarding operations of each node in a fully decentralized and localized manner. Each node overhears the channel, monitors the behaviour of its neighbours, and discovers consistent misbehaviour as indications of attacks. Moreover, local neighbouring nodes collaborate with each other to improve the monitoring accuracy.

1) Monitor Routing Behaviour:

Our basic idea is to overhear the channel and *cross-check* the routing messages announced by different nodes. This can be applied to any distributed and

deterministic routing protocol. In such protocols, the routing activity of a node is a three-step process: 1) receiving routing updates from neighbouring nodes as inputs to a routing algorithm

2) Executing the routing algorithm; and 3) announcing the output of the routing algorithm as its own routing updates. The monitoring task is to verify whether the routing algorithm executed by a node follows the protocol specifications. In other words, the trustworthiness of a routing message, as the output of the routing algorithm, can be examined when the monitoring node knows the input to the algorithm, since the algorithm itself is publicly known and deterministic. This idea can be illustrated with the context of AODV. An AODV node cannot acquire enough information about the routing algorithms. The key reason is that the *next hop* information is missing in the AODV routing messages. Thus, when a node announces a routing update, its neighbours have no hint about which node is the next hop in the route and, hence, cannot judge on its input to the routing algorithm, i.e., the original routing update on which its routing computation is based .In order to permit the cross-checking of routing updates, two modifications to AODV need to be done. First, add one more field, *next_hop*, in the RREP packet. Similarly, add one more field, *previous_hop*, in the RREQ packet. This way, each node explicitly claims its next hop in a route when it advertises routing updates. Second, each node keeps track of the routing updates previously announced by its neighbours. Essentially, each node maintains part of the routing tables of its neighbours.

2) Monitor Packet Forwarding Behaviour:

Each SCAN node also monitors the packet forwarding activity of its neighbours. This is achieved by overhearing the channel and comparing ongoing data transmission with previously recorded routing messages. The currently focus has been on three kinds of forwarding misbehaviour, namely, packet drop, packet duplication, and network-layer packet jamming, and develop simple algorithms to detect each of them. Packet drop means that a node drops the packets that it is supposed to forward for its neighbours; packet duplication means that a node duplicates the packets that it has already forwarded; and packet jamming means that a node sends too many packets and occupies a significant portion of the channel bandwidth In ,The packet drop detection algorithm in which the sender explicitly lists the route in the data packet header. It cannot be directly applied in the AODV context, because when a node receives a packet, its neighbours do not know to which node it should forward the packet, thus, cannot tell whether it forward the packet in the correct manner. Fortunately, our modification to the AODV protocol, described in the previous section, enables the detection of packet drop, because each node keeps track of the route entries announced by its neighbours, which explicitly lists the *next_hop* field. Specifically, each SCAN node records the headers of the recent packets it has overheard. If one node overhears that the bandwidth consumed by

duplicate packets from its neighbour exceeds the threshold *Duplicate_Bandwidth*, or the bandwidth consumed by packets originated from its neighbour exceeds the threshold *Sending Bandwidth*, it also considers these events as packet forwarding misbehaviour.

VI. SIMULATION EVALUATION

In this section, we evaluate the performance of SCAN through extensive simulations, the goal of which is to answer the following questions

In the simulations, the following metrics are observed: 1) *miss detection ratio*, which is the chance that SCAN fails to convict and isolate a malicious node; 2) *false accusation ratio*, which is the chance that SCAN incorrectly convicts and isolates a legitimate node; 3) *packet delivery ratio*, which is the percentage of packets that are successfully delivered to the receiver nodes; and 4) *communication overhead*, which is the total number of packets sent by SCAN in order to achieve its goal.

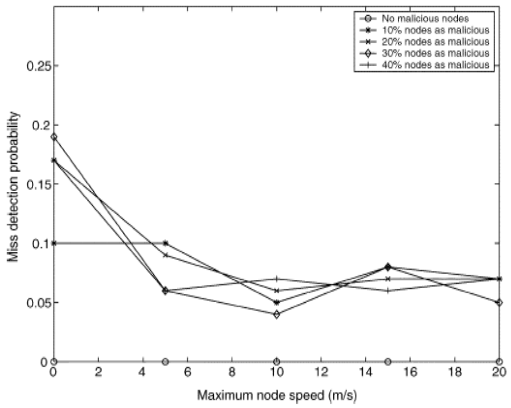


Fig. 1. Miss detection probability versus mobility.

the miss detection ratio obtained by considering only the set of active malicious nodes, instead of all prechosen malicious nodes. The false accusation ratio is obtained in a similar way over the set of active legitimate nodes

B. Monitoring and Detection

The detection performance of the collaborative monitoring mechanism in SCAN in terms of miss detection and false accusation ratios .Fig 1 shows the miss detection ratio as the node mobility speed changes. Ratio is the highest in a static network, regardless of the number of malicious nodes. The miss detection

ratio drops considerably when nodes start to move, and remains stable at 4%–8% when the speed further increases. SCAN fails to convict a malicious node mainly because it resides in a sparsely occupied region. In a static network, if a malicious node happens to stay in a sparsely occupied region, its neighbours always have no chance to convict it. On the contrary, in a mobile network, the mobility increases the chance that other nodes roam into this region or the malicious node itself moves into another densely occupied region. The impact of node mobility on the false accusation ratio is presented in Fig. 7 the false accusation ratio continues to increase as nodes move faster. When the maximum speed is 20 m/s, the false accusation ratio is around 5% 10%. The reason is that higher mobility makes nodes more "memory less. Fig. 6 and 7 also illustrate the impact of the number of malicious nodes on the detection performance. In both cases, even if the number of malicious nodes increases dramatically from 0% to 40% of the network population, it does not exhibit evident impact on the detection performance.

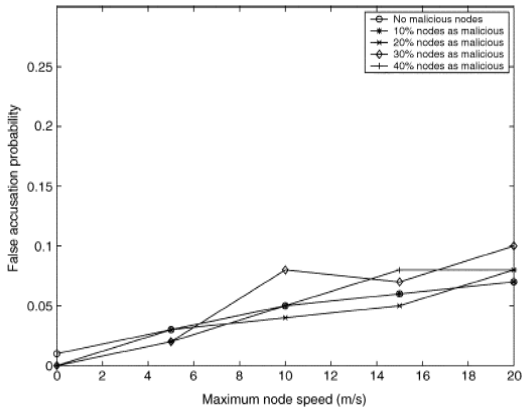


Fig. 2. False accusation probability versus mobility

C. Packet Delivery Ratio

Fig. 8 shows the improvement on the packet delivery ratio in a SCAN-protected network. In these simulations, 30% of the nodes are set as malicious nodes. From the figure that SCANS increases the packet delivery ratio by a factor up to 150% even if 30% of nodes are malicious. In an ad hoc network without any security protection, the packet delivery ratio can be as low as 30%, even if the network is lightly loaded. Another observation from Fig. 3 is that even in a SCAN

protected and light-loaded network, the packet delivery ratio is not 100%.

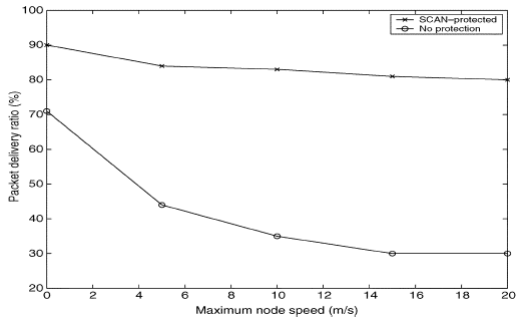


Fig. 3. Packet delivery ratio versus mobility.

VII. CONCLUSION

Here in this paper, the study of various attack has been done. Mobile ad hoc network is vulnerable to various attack Some of the vulnerabilities of the Mobile ad hoc network has been studied. This paper explores a novel self-organized approach to securing such networks. To this end, SCAN technology is presented, a network-layer security solution that protects routing and forwarding operations in a unified framework. SCAN exploits localized collaboration to detect and react to security threats. All nodes in a local neighbourhood collaboratively monitor each other and sustain each other, and no single node is superior to the others. Both analysis and simulations results have confirmed the effectiveness and efficiency of SCAN in protecting the network layer in mobile ad hoc networks

REFRENCES

[1] S. Marti, T. Giuli, K. Lai, and M. Baker, “Mitigating routing Misbehaviour in mobile ad hoc networks,” in *Proc. ACM MobiCom*, 2000, pp. 255–265.

[2] J. Hubaux, L. Buttyan, and S. Capkun, “The quest for Security in mobile ad hoc networks,” in *Proc. ACM MobiHoc*, 2001, pp. 146–155.

[3] J. Kong, P. Zerfos, H. Luo, S. Lu, and L. Zhang, “Providing Robust and ubiquitous security support for MANET,” in *Proc. IEEE ICNP*, 2001, pp. 251–260.

[4] Y. Hu, A. Perrig, and D. Johnson, "Ariadne: A secure on-Demand routing protocol for ad hoc networks," in *Proc. ACM MobiCom*, 2002, pp. 12–23.

[5] Y.C. Hu, A. Perrig, and D.B. Johnson, "Rushing Attacks And Defense in Wireless Ad Hoc Network Routing Protocols", Proceedings *of ACM WiSe 2003*, San Diego, CA, Sep. 2003

[6] M. Zapata and N. Asokan, "Securing ad hoc routing Protocols," in *Proc. ACM WiSe*, 2002, pp. 1–10.

[7] C. Perkins and E. Royer, "Ad hoc on-demand distance vector routing," in *Proc. IEEE WMCSA*, 1999, pp. 90–100.

[8] C. Perkins, E. Royer, and S. Das, "Ad hoc on demand Distance vector (AODV) Routing," Internet Draft, draft-ietf manet-aodv-10.txt, 2002.

[9] Y. Hu, A. Perrig, and D. Johnson, "Packet leashes: A Defense against wormhole attacks in wireless ad hoc networks," in *Proc. IEEE INFOCOM*, 2003, pp. 1976–1986.

[10] L. Buttyan and J. Hubaux, "Stimulating cooperation in self-organizing mobile ad hoc networks," *ACM/Kluwer Mobile Netw. Applicat.*, vol. 8, no. 5, pp. 579–592, Oct. 2003.

IT 007

Hand Gesture Recognition System

Authors

Prof. S.B.Chaudhari
P.D.E.A's College Of Engineering Manjari, Pune University, Pune, Maharashtra, India.

Mr. Krushna Belerao (B.E.Computer Science),
P.D.E.A's College Of Engineering Manjari, Pune University, Pune, Maharashtra, India.
kriushnabelerao@gmail.com

Mr. Pratik Gawali (B.E. Computer Science),
P.D.E.A's College Of Engineering Manjari, Pune University, Pune, Maharashtra, India.
princeofkaviraj@gmail.com

Abstract

In this paper we are providing a implementation details about simulated solution of Hand Gesture Recognition. This System will design and build a man-machine interface using a video camera to interpret the American one-handed sign language alphabet gestures. The keyboard and mouse are currently the main interfaces between man and computer. Humans communicate mainly by vision and sound, therefore, a man-machine interface would be more intuitive if it made greater use of vision and audio recognition. Advantage is that the user not only can communicate from a distance, but need have no physical contact with the computer. However, unlike audio commands, a visual system would be preferable in noisy environments or in situations where sound would cause a disturbance. The way humans interact with computers is constantly evolving, with the general purpose being to increase the efficiency and effectiveness by which interactive tasks are completed. Real-time, static hand gesture recognition affords users the ability to interact with computers in more natural and intuitive ways.

Keywords

Gesture Recognition, Image Processing, Human computer Interaction, Computer Vision.

1.

Introduction

With the development of information technology in our society, we can expect that computer systems to a larger extent will be embedded into our environment. These environments will impose needs for new types of human computer-interaction, with interfaces that are natural and easy to use. The user interface (UI) of the personal computer has evolved from a text-based command line to a graphical interface with keyboard and mouse inputs. However, they are inconvenient and unnatural. The use of hand gestures provides an attractive alternative to these cumbersome interface devices for human-computer interaction (HCI). User's generally use hand gestures for expression of their feelings and notifications of their thoughts. In particular, visual interpretation of hand gestures can help in achieving the ease and naturalness desired for HCI. Vision has the potential of carrying a wealth of information in a nonintrusive manner and at a low cost, therefore it constitutes a very attractive sensing modality for developing hand gestures recognition. Recent researches in computer vision have established the importance of gesture recognition systems for the purpose of human computer interaction.

Features for gesture Recognition

Selecting features is crucial to gesture recognition, since hand gestures are very rich in shape variation, motion and textures. For static hand posture recognition, although it is possible to recognize hand posture by extracting some geometric features such as fingertips, finger directions and hand contours, such features are not always available and reliable due to self-occlusion and lighting conditions. There are also many other non-geometric features such as color, Silhouette and textures, however, they are inadequate in recognition. Since it is not easy to specify features explicitly, the whole image or transformed image is taken as the input and features are selected implicitly and automatically by the recognizer. Vision-based interaction is a challenging interdisciplinary research area, which involves computer Vision and graphics, image processing, machine learning, bio-informatics, and psychology. To make a successful working system, there are some requirements which the system should have:

(a) ***Robustness:*** In the real-world, visual information could be very rich, noisy, and incomplete, due to changing illumination, clutter and dynamic backgrounds, occlusion, etc. Vision-based systems should be user independent and robust against all these factors.

(b) ***Computational Efficiency****:* Generally, Vision based interaction often requires real-time systems. The vision and learning techniques/algorithms used in Vision-based interaction should be effective as well as cost efficient.

(c) ***User's Tolerance****:* The malfunctions or mistakes of Vision-based interaction should be tolerated. When a mistake is made, it should not incur much loss. Users can be asked to repeat some actions, instead of letting the computer make more wrong decisions.

(d) ***Scalability:*** The Vision-based interaction system should be easily adapted to different scales of applications. For e.g. the core of Vision-based interaction should be the same for desktop environments, Sign Language Recognition, robot navigation and also for VE.

Related Work

Recognizing gestures is a complex task which involves many aspects such as motion modeling, motion analysis, pattern recognition and machine learning, even psycholinguistic studies. The enormous potential for sophisticated and natural human computer interaction using gesture has motivated work as long ago as 1980 with systems such as Bolt's seminal "Put-That-There".

Whilst "Put-That-There" used a data glove as input and stylus drawing, video has been used in more recent systems. Video-based American Sign Language recognition system is a worthwhile and impressive achievement. A vision-based system able to recognize 26 gestures in real time to manipulate windows and objects within a graphical interface was developed by Ng *et al.* Describe a system that recognizes hand gestures through the detection of the bending of the hand's five fingers, based on image-property analysis. *A r*eal-time hand gesture recognition system using skin color segmentation and multiple-feature based template-matching techniques. In their method, the three largest skin-like regions are segmented from the input images by skin color segmentation technique from color space and they are compared for feature based template

matching using a combination of two features: correlation coefficient and minimum (Manhattan distance) distance qualifier. These Gesture commands are being through Software Platform for Agent and Knowledge Management (SPAK) and their actions are being accomplished according to user's predefined action for that gesture.

2. Gestures Taxonomies

Several taxonomies have been suggested in the literature that deals with psychological aspects of gestures. They vary from author to author .This distinguishes "autonomous gestures" which occurs independently of speech from "gesticulation" that occurs in association of speech. The taxonomy that seems most appropriate for HCI purposes (Figure 1) .

Figure 1. Taxonomies for Hand Gesture

Gesture Recognition Architecture

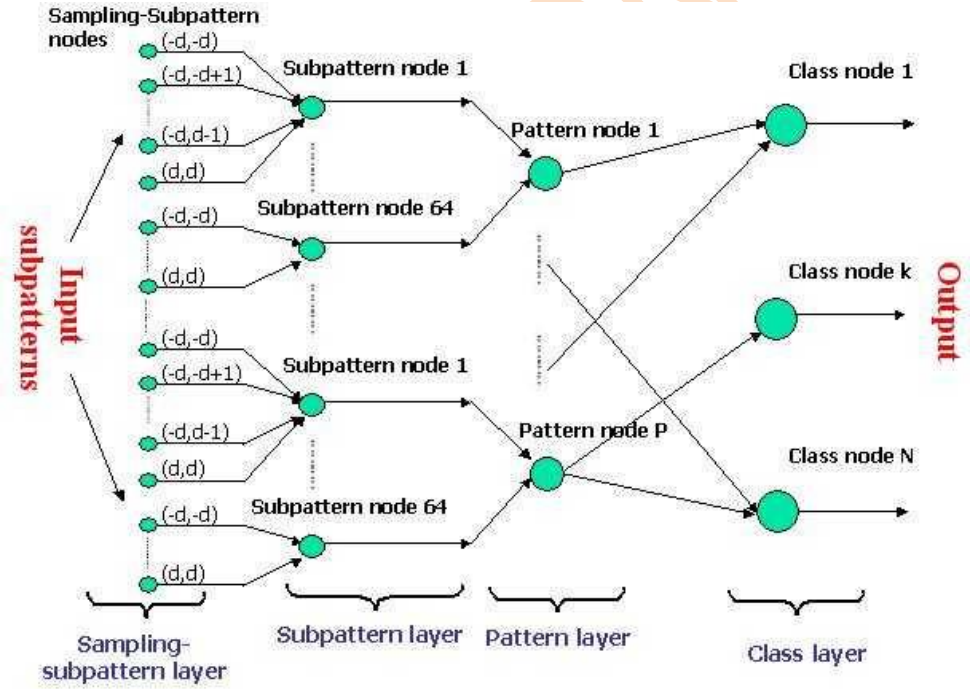


Figure 2. Architecture for gesture Recognition

A gesture recognition consists of four layers, excluding the input layer, as show in Fig.2. Subpatterns of an input pattern are presented to the sampling-sub pattern layer. Sampling-subpattern nodes take care of the deformation, i.e., shift, noise, or size, of the input pattern. Each subpattern node summarizes the measure of similarity measure between the input pattern and the stored pattern. Subpattern node is responsible for the match between an input nominal sub pattern and the stored pattern. However, for tolerating possible deformation of the input pattern, we have to consider the neighboring subpatterns of an input subpattern. Suppose we allow a deformation of up to ± d (d is a positive integer) pixels in either X or Y directions. We have to consider all the neighboring sub patterns within the distance d in order to detect a possible deformation. Each neighboring subpattern is taken care of in a sampling-subpattern node. Therefore, a subpattern node may receive the output of up to (2d+1) 2 samplingpattern nodes. Each subpattern node stores. A sampling sub pattern node weight W, which is shared by all its sampling-subpattern nodes. A sampling-subpattern node computer based on the input pattern and its node weight, a value and outputs the value to the associated subpattern node.

Scope Of System

The application to be developed is Hand Gesture Recognition System (HGRS). Hand Gesture Recognition System project involves generating characters, words from hand gesture. Through this project we achieve the following objectives:

Develop a user-friendly system that can detect the characters and words from the captured frames and recognize its patterns.
Recognition of ASL (American Sign Language) using hand gestures.
To overcome challenges like dynamic background and skin color detection.
To recognize the alphabets and words using hand gestures.
Providing the easy way of interacting with computer.

Future Scope

The project can be extended to make sentences by capturing the video of hand gesture. We will also convert the recognized character to the voice. Through this application dumb peoples can be make announcement.

Technology and Features

We are going to develop this project in software platform .NET (2010) with help of AForge.NET libraries. The framework's API support for:

Computer vision, Image, Video processing; comprehensive image filter library; Neural networks;

Genetic programming;
Fuzzy logic.
Machine Learning;
Libraries for a select set of robotics kits.

Coclusion

We are developing a gesture recognition system that is proved to be robust for ASL gestures. The system is fully automatic and it works in real-time, static background. The advantage of the system lies in the ease of its use. The users do not need to wear a glove, neither is there need for a uniform background. Experiments on a single hand database have been carried out.

We plan to extend our system into tracking sentences and convert it into voice. Currently, our tracking method is limited to single characters. Focus will also be given to further improve our system and the use of a lager hand database to test system and recognition. It is efficient as long as the data sets are small and not further improvement is expected. Our major goal was speed and the avoidance of special hardware.

Acknowledgement

Our thanks to the experts who have contributed towards development of the stealthy attack and its simulated solution. We would like to thank everyone, just everyone!

References

July-December 2009, Volume 2, No. 2, pp. 405-410
A Review of the Hand Gesture Recignition System

G. R. S. Murthy & R. S. Jadon

World Academy of Science, Engineering and Technology 50 2009
Real-Time Hand Tracking and Gesture Recognition System Using Neural Networks

Tin Hninn Hninn Maung

GVIP 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt
Hand Gesture Recognition Using Fuzzy Neural Network
Nguyen Dang Binh, Toshiaki Ejima

A Fast Algorithm for vision based Hand Gesture Recognition for Robot Control 2005
Asanterabi Malima, Erol Özgür, and Müjdat Çetin

R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, 2nd edition, 2002.

World Academy of Science, Engineering and Technology 60 2009
Hand Gesture Recognition Based On Combined feature extraction

Mahmoud Elmezain, Ayoub Al-Hamadi, and Bernd Michaelis

Institute for Electronics, Signal Processing and Communications (IESK)

Otto-von-Guericke-University Magdeburg

A parallel Algorithm for Real time Hand Gesture Recognition. 2010
Varun Ramesh , Hamilton high School, Chandler, Arizone

Konstantinos G. Derpanis. "A Review of Vision-Based Hand Gestures" (2004).

Sushmita Mitra, and Tinku Acharya, "Gesture Recognition: A Survey",

IEEE Transactions on Systems, Man and Cybernetics–Part C: Applications and Reviews, (2007).

IT 008

CONTEXT BASED ENGLISH TO MARATHI LANGUAGE TRANSLATOR

Sunita Vijay Patil

Designation- B.E Student (Comp)

City- Pune

State- Maharashtra

Country- India

Email-id- sunpatil.patil@gmail.com

Contact-8237370189

Kalpana Satish Musne

Designation- B.E Student (Comp)

City- Pune

State- Maharashtra

Country- India

Email-id- kalpana.musne@gmail.com

Contact-9970329875

Abstract

Context Based English to Marathi Translator - converting Simple English affirmative sentence to Marathi. In this project we are converting the simple English affirmative sentences to Marathi sentences. This is basically a machine translation. We have chosen the transfer based approach which is the thin line between the semantic and the direct approach. For that we have designed the parser which helps us to map the English sentence binding to the rules and then getting converted into target language. In this project we are going through various processes such as morphological analysis, part of speech, local

word grouping, and all this using perspective for converting the meaning of simple affirmative English sentence into corresponding Marathi sentence. In this system we have included a speech output also. After the user gives an input the system will process it & as a result the system will produce output in two different ways. First it will show the text output to the user in Marathi language & second it will produce a sound of the generated output text. The system will use Phonetic Conversion for speech. The SAPI will be used for this purpose. What the system do-: After the user input the system create token and parse the input and Check whether syntactically correct, Does morphological analysis and local word grouping, Convert to the equivalent Marathi and then correct structurally. Help people to communicate comfortably. A successful solution would be to implement the conversion of one language to another to help communication between the two linguistically different backgrounds. Keeping the time factor into consideration we are not including the tagged questions, interrogative sentences, and interrogative negatives sentences.

Keywords- *affirmative sentences, lexical parser, morphological analysis, NLP, SVO.*

I. INTRODUCTION

The purpose of this document is to collect, analyze and define high-level needs and features of

the English to Marathi language Translator. The detailed analysis of the requirements of the system

is done and details of how the systems fulfils these requirements are detailed in the use-case

specifications. The scope is limited to the affirmative sentence. The system objective is to provide

machine translation of the English sentence. The system of MT of so tedious because of the different

background in the language the Eng sentence is in the free word format also the morphological

analysis is the another Herculean task so keeping the time consideration we have reduce out scope to

the affirmative sentence. English to Marathi language Translator (EMLT) is a field of computer

science and linguistics concerned with the interactions between computers and human (natural)

language. EMLT systems convert information from computer databases into readable human language.

Natural systems convert samples of human language into more formal representations such as parse

trees or first-order logic structures that are easier for computer programs to manipulate. Many

problems within EMLT apply to both generation and understanding; for example, a computer must be

able to model morphology (the structure of words) in order to understand an English sentence, and a

model of morphology is also needed for producing a grammatically correct English sentence. To

implement an English to Marathi language converter with facility of voice playback of the generated

output. After the user input the system create token and parse the input and Check whether

syntactically correct. Morphological structure used to concatenate the grammar to form words.

Innovativeness: It may reduce the Ambiguity as the context based approach is used in this software. The Syntactic-Semantic approach will be used & the algorithmic structural transformation can be used to apply for any other formal languages. This algorithmic structure we can use in any other translator to convert a formal language.

II. TECHNOLOGIES USED IN TRANSLATOR

There are some technologies used in the Context Based Language Translator (CBLT) which has different approaches & methods.

A. SAPI (Speech Application Programming Interface)

The Speech Application Programming Interface or SAPI is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. To date, a number of versions of the API have been released, which have shipped either as part of a Speech SDK, or as part of the Windows OS itself. Applications that use SAPI include Microsoft Office, Microsoft Agent and Server. In general all versions of the API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages. In addition, it is possible for a 3rd-party company to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with SAPI. In principle, as long as these engines conform to the defined interfaces they can be used instead of the Microsoft-supplied engines. In general the Speech API is a freely-redistributable component which can be shipped with any Windows application that wishes to use speech technology. Many versions (although not all) of the speech recognition and synthesis engines are also freely redistributable.

B. NET BEANS

The NetBeans IDE is written in Java and can run anywhere a compatible JVM is installed, including Windows, Mac OS, Linux, and Solaris. A JDK is required for Java development functionality, but is not required for development in other programming languages. The NetBeans platform allows applications to be developed from a set of modular software components called *modules*. Applications based on the NetBeans platform The NetBeans Platform is a reusable framework for simplifying the development of Java Swing desktop applications. The NetBeans IDE bundle for Java SE contains what is needed to start developing NetBeans plugins and NetBeans Platform based applications; no additional SDK is required.

The platform offers reusable services common to desktop applications, allowing developers to focus on the logic specific to their application. Among the features of the platform are:

- User interface management (e.g. menus and toolbars)
- User settings management
- Storage management (saving and loading any kind of data)
- Window management
- Wizard framework (supports step-by-step dialogs)

- NetBeans Visual Library
- Integrated Development Tools

C. DATABASE / DATA LIBRARY

A data library refers to both the content and the services that foster use of collections of numeric, audio-visual, textual or geospatial data sets for secondary use in research. (See below to view definition from the *Online Dictionary for Library and Information Science*.) A data library is normally part of a larger institution (academic, corporate, scientific, medical, governmental, etc.) established to serve the data users of that organisation. The data library tends to house local data collections and provides access to them through various means (CD-/DVD-ROMs or central server for download). A data library may also maintain subscriptions to licensed data resources for its users to access. Whether a data library is also considered a data archive may depend on the extent of unique holdings in the collection, whether long-term preservation services are offered, and whether it serves a broader community. the Association of Research Libraries (ARL) published SPEC Kit 263: *Numeric Data Products and Services*, presenting results from a survey of ARL member institutions involved in collecting and providing services for numeric data resources.

D. PROCESS BUILDER

public final class ProcessBuilder
extends Object

This class is used to create operating system processes.

Each ProcessBuilder instance manages a collection of process attributes. The start() method creates a new Process instance with those attributes. The start() method can be invoked repeatedly from the same instance to create new sub processes with identical or related attributes. Each process builder manages these process attributes:

- A *command*, a list of strings which signifies the external program file to be invoked and its arguments, if any. Which string lists represent a valid operating system command is system-dependent. For example, it is common for each conceptual argument to be an element in this list, but there are operating systems where programs are expected to tokenize command line strings themselves - on such a system a Java implementation might require commands to contain exactly two elements.

- An *environment*, which is a system-dependent mapping from *variables* to *values*. The initial value is a copy of the environment of the current process (see System.getenv()).
- A *working directory*. The default value is the current working directory of the current process, usually the directory named by the system property user.dir.
- A *redirectErrorStream* property. Initially, this property is false, meaning that the standard output and error output of a subprocess are sent to two separate streams, which can be accessed using the Process.getInputStream() and Process.getErrorStream() methods. If the value is set to true, the standard error is merged with the standard output. This makes it easier to correlate error messages with the corresponding output. In this case, the merged data can be read from the stream returned by Process.getInputStream(), while reading from the stream returned by Process.getErrorStream() will get an immediate end of file.

Modifying a process builder's attributes will affect processes subsequently started by that object's start() method, but will never affect previously started processes or the Java process itself. Most error checking is performed by the start() method. It is possible to modify the state of an object so that start() will fail. For example, setting the command attribute to an empty list will not throw an exception unless start() is invoked.

E. GUI - AWT & SWING

1. AWT

The Abstract Window Toolkit (AWT) is Java's original platform-independent windowing, graphics, and user-interface widget toolkit. The AWT is now part of the Java Foundation Classes (JFC) — the standard API for providing a graphical user interface (GUI) for a Java program. AWT is also the GUI toolkit for a number of Java ME profiles. For example, Connected Device Configuration profiles require Java runtimes on mobile telephones to support AWT.

The AWT provides two levels of APIs:

- A general interface between Java and the native system, used for windowing, events, and layout managers. This API is at the core of Java GUI programming and is also used by Swing and Java 2D. It contains:
  - The interface between the native windowing system and the Java application;
  - The core of the GUI event subsystem;

- Several layout managers;
- The interface to input devices such as mouse and keyboard; and
- A java.awt.datatransfer package for use with the Clipboard and Drag and Drop.

- A basic set of GUI widgets such as buttons, text boxes, and menus. It also provides the AWT Native Interface, which enables rendering libraries compiled to native code to draw directly to an AWT Canvas object drawing surface.

AWT also makes some higher level functionality available to applications, such as:

- Access to the system tray on supporting systems; and
- The ability to launch some desktop applications such as web browsers and email clients from a Java application.

Neither AWT nor Swing are inherently thread safe. Therefore, code that updates the GUI or processes events should execute on the Event dispatching thread. Failure to do so may result in a deadlock or race condition. To address this problem, a utility class called SwingWorker allows applications to perform time-consuming tasks following user-interaction events in the event dispatching thread.

2. SWING

Swing is the primary Java GUI widget toolkit. It is part of Oracle's Java Foundation Classes (JFC) — an API for providing a graphical user interface (GUI) for Java programs. Swing was developed to provide a more sophisticated set of GUI components than the earlier Abstract Window Toolkit. Swing provides a native look and feel that emulates the look and feel of several platforms, and also supports a pluggable look and feel that allows applications to have a look and feel unrelated to the underlying platform. It has more powerful and flexible components than AWT. In addition to familiar components such as buttons, check box and labels, Swing provides several advanced components such as tabbed panel, scroll panes, trees, tables and lists. Unlike AWT components, Swing components are not implemented by platform-specific code. Instead they are written entirely in Java and therefore are platform-independent. The term "lightweight" is used to describe such an element.

III. SYSTEM ARCHITECTURE

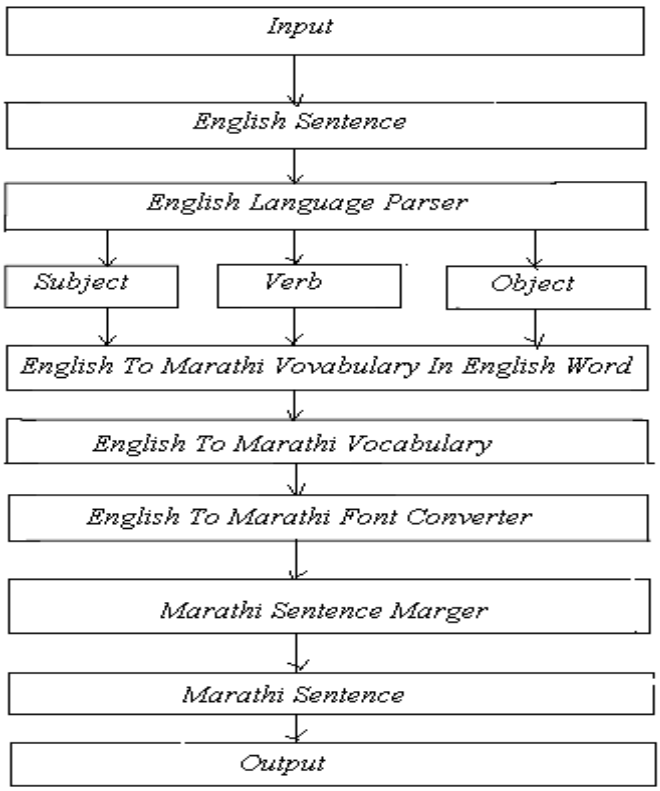


Fig. 1 Architecture of a typical CBLT system

Features-:

1. Input should be an English Tex.
2. To parse English Sentence we use a Lexical Parser.
3. English to Marathi Vocabulary in English word.
4. English-Marathi Vocabulary
5. English-Marathi to Marathi font Converter.
6. Marathi Sentence Merger.
7. Output would be an Marathi Sentence.

Modules of the System Architecture-:

Module 1: Lexical Parser

Creation of parser to parse English sentence lexically and provide a lexical tree structure or data

Structure.

Module 2: Semantic Mapper

To map the English semantic word with Marathi semantic word.

Module 3: ITranslation

To convert English-Marathi word to Marathi(font).

Module 4: Composer

To get the Marathi word and form the sentence as per the Marathi Grammar.

Algorithms for Modules-:

1. LEXICAL PARSER

Step 1: Tokenize the sentence into various tokens i.e. token list
Step 2: To find the relationship between tokens we are using dependency grammar and binary
relation for our English Lang token list acts as an input to semantic class to represent the
semantic standard
Step 3: Semantic class generates a tree we have a class tree transform which will create a tree
Step 4: Sentence is splitted into words that are nouns, verbs etc.

Example:
The Sentence: Bell based in Los Angeles, makes and Distributes electronic, computer and building products.
These are broken into tokens

Noun(nsubj)	Token1	Bell

Participial modifier (partmod)	Token2	Based
Preposition(prep)	Token3	In
Noun(nn)	Token4	Los
Noun(nn)	Token5	Angeles
Verb	Token6	Makes
Conjunction(conj)	Token7	And
Verb	Token8	Distributes
Adjective(amod)	Token9	Electronic
Adjective(amod)	Token10	Computer
Conjunction(conj)	Token11	And
Adjective(amod)	Token12	Building
Directobject(dobj)	Token13	Products

Table shows how sentence is broken into tokens

2. **SEMANTIC MAPPER**

Step 1: The output from the first Module i.e. Lexical Parser acts as input to the Semantic Mapper
Step 2: The tokens generated from the first Module is stored in data structure i.e. collection.
These tokens has grammatical relations which are represented with various symbols e.g.
conj, nn, nsubj, det, dobj etc
Step 3: Look up in Marathi dictionary we are matching English semantic word with the
dictionary Marathi word. This matching is not only word by word but it will be
semantic(meaningful) matching based on the relationship been established

Step 4: After matching the selected words from the Marathi dictionary are kept as another data
structure
Step 5: Identify the relationships among the various Marathi words from these data structures.

Example:
Different rules are considered for Mapping

EQUALITY RULE-:

English word directly mapping to Marathi word
e.g. ES: A man eats vegetables
SS: Manushya Bhajya Khato
(A) (man) (vegetables) (eats)

SYNONYMS RULE (word having same meaning)

English words mapping to Marathi word
e.g. ES: He is a good/fine/excellent man
SS: Sabhya
ANTONYMS RULE (word having opposite meaning)

English word not directly mapping to Marathi word
e.g. ES: He is not good man
He is a bad man
SS: Asabhya

3. ITRANSLATOR

The English Marathi words are converted to Marathi words i.e. in Marathi font using localization and internationalization.

4. COMPOSER

The English sentence after parsing after mapped with the bilingual dictionary generates the English Marathi words then using the ITranslator the English Marathi words are converted to Marathi words i.e. in Marathi words that acts as an input to this Module it composes the words into a sentence.

IV. CONCLUSION

The application performs a simple language translator in the Java platform for simple Affirmative sentences, using Context based concept. The system or MT(machine translation) engine with a syntactic-semantic approach will be formed. Algorithmic structural transformation can be applied to other languages as well. We just have to analyze grammatical structure of target language and transform the tree accordingly. This system will play an important role in translation.

REFERENCES

[1]. About FnTBL POS Tagger. Available
http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html

[2]. Dan Melamed, Statistical Machine Translation by
Parsing, Proceedings of ACL, 2004.

[3]. Deepa Gupta, Niladri Chatterjee – A Morpho – Syntax Based Adaptation and Retrieval Scheme for English to Hindi, Department of Mathematics, IIT 0 Delhi

[4]. Deepa Gupta, Niladri Chatterjee – Study of Divergence for Example Based English-Hindi Machine Translation. Department of Mathematics, IIT 0 Delhi.

[5]. Dekang Lin. An Information – Theoretic Definition of Similarity. Department of CS, University of Manitoba, 1998.

[6]. Doug Arnold, Lorna Balkan, Siety Meijer, R. Lee Humpreys, Lousia Sadler.

Machine Translation: An Introductory Guide. University of Essex, 2002.

[7]. Goyal P. Mital Manav R. Mukerjee A. Shukla P. and Vikram K. A bilingual parser for Hindi, English and Code Switching structures, IIT – Kanpur.

[8]. Jurafsky Daniel, Martin James H. - Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, 2000.

[9]. Kenji Imamura, Hideo Okuma, Eiichiro Sumita, Practical Approach to Syntax-based Statistical Machine Translation, Proceedings of MTSUMMIT X, pages

[10]. Kevin Knight, - A Statistical MT Tutorial Workbook, prepared in connection with the JHU summer workshop, April 1999.

[11]. Niladri Chatterjee, A Statistical Approach for Similarity Measurement Between Sentences for EBMT, Department of Mathematics, IIT – Delhi.

[12]. Phrasal Verb and Multi-word Verbs Available: http://grammar.englishclub.com/verbs-phrasal-verbs.htm

IT 009

DECISION TREE BASED CLASSIFICATION: A DATA MINING APPROACH

Ms. Rita R. Kamble
reetakamble@ in.com
Ph. 9920295977

Abstract—Decision tree is mainly used for model classification and prediction. ID3 algorithm is the most widely used algorithm in the decision tree. ID3 is famous for high classifying speed easy, strong learning ability and easy construction. Through illustrating on the basic ideas of decision tree in data mining, in this paper, the shortcoming of ID3's is discussed, and then a new decision tree algorithm combining ID3 and Association Function (AF) is presented. The experiment results show that the proposed algorithm can get more effective rules than ID3.

Key Terms—data mining, decision tree, ID3, association function (AF).

INTRODUCTION

With the rising of data mining,

decision tree plays an important role in the process of data mining and data analysis.

With the development of computer technology and network technology, the degree of informationization is getting higher and higher, people's ability of using information technology to collect and produce data is substantially enhanced. Data mining is a process to extract information and knowledge from a large number of incomplete, noisy, fuzzy and random data. In these data, the information advance, but potentially useful. At present, the decision tree has become an important data mining method. The basic learning approach of decision tree is greedy algorithm, which use the recursive top-down approach of decision tree structure.

Quin lan in 1979 put forward a well-known ID3 algorithm, which is the most widely used algorithm in decision tree. But that algorithm has a defect of tending to use attributes with many values. Aiming at the shortcomings of the ID3 algorithm, in the paper, an association function is introduced to improve ID3 algorithm. The result of experiment shows that the presented algorithm is effective than ID3.

I. I

D3 ALGORITHM

Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree. Thus, we can select the attribute with the highest information gain as the test attribute of current node. In this way, the information needed to classify the training sample subset obtained from later on partitioning will be the smallest. That is to say, the use of this property to partition the sample set contained in current node will make the mixture degree of different types for all generated sample subsets reduce to a minimum. Therefore, the use of such an information theory approach will effectively reduce the required dividing number of object classification.

Set S is set including s number of data samples whose type attribute can take m potential different values corresponding to m different types of Ci (1,2,3, …, m). Assume that s is the sample number of Ci. Then, the required amount of information to classify a given data is

I (S1, S2, …., Sm) = -∑ Pi log Pi

Where P is the probability that any subset of data samples belonging to categories Ci.

Suppose that A is a property which has v different values. Using the property of A, S can be divided into v number of subsets in which S contains data samples whose attribute A are equal aj in S set. If property A is selected as the property for test, that is, used to make partitions for current sample set, suppose that S is a sample set of type Ci in subset Si ,the required information entropy is

$E(A) = \sum_{j}^{v} S_{ij} + ... + S_{mj} / S * I(S_{ij} + ... + S_{mj})$ Such use of property A on the current branch node corresponding set partitioning samples obtained information gain is:

Gain (A) = I ($S_1, S_2, ...S_m$) - E (A)

ID3 algorithm traverses possible decision-making space using top-down greedy search strategy, and never trace back and reconsider previous selections. Information gain is exactly the metrics for selecting the best attribute in each

step of the algorithm for generating a decision tree according to a given data sets.

Input: training samples, each attribute taking discrete value, a candidate attribute set available for induction is attribute_list.

Output: a decision tree.

Deal flow:

1)
Create a node N;

2)
If (All samples are in same class) Return node as leaf with classname;
3)
If (attribute list is empty)
Return node as leaf node labelled with most common class;

4)
Select list attribute i.e attributes having highest information gain
5)
Label node N with test attribute
6)
for each known value of a; of test attribute, grow branches from node N for the condition test attribute =a;
7)
Let Si be set of samples for which test attribute = ai;
8)
If (Si is empty) then attach the leaf labelled with most common class in sample
9)
Else attach the node returned by generate_decision_tree (Si,attribute_list_test_attribute).

This is a greedy algorithm which use recursive manner of top-down, divide and conquer to construct a decision tree. The termination condition of recursion is : all samples within a node are of the same category. If no attribute can be used to divide current sample set, then voting principle is used to make it a Compulsory leaf node, and mark it with the category of having the most number of sample types. If no sample satisfies the *a* , then a leaf node is created, and mark it with the category of having the most number of sample condition of test-attribute = *ai* then a leaf node is created, and mark it with the category of having the most number of sample types.

THE SHORTCOMING OF ID3 ALGORITHM :

The principle of selecting attribute A as test attribute for ID3 is to make E (A) of attribute A, the smallest. Study suggest that there exists a problem with this method, this means that it often biased to select attributes with more taken values, however, which are not necessarily the best attributes. In other words, it is not so important in real situation for those attributes selected by ID3 algorithm to be judged firstly according to make value of entropy minimal. Besides, ID3 algorithm selects attributes in terms of information entropy which is computed based on probabilities, while probability method is only suitable for solving stochastic problems. Aiming at these shortcomings for ID3 algorithm, some improvements on ID3 algorithm are made and a improved decision tree algorithm is presented.

II.

THE IMPROVED OF ID3 ALGORITHM

To overcome the shortcoming stated above, attribute related method is firstly applied to computer the importance of each attribute. Then, information gain is combined with attribute importance, and it is used as a new standard of attribute selection to construct decision tree. The conventional methods for computing attribute importance are sensitivity analysis (SA), information entropy based joint information entropy method(MI), Separation Method(SCM), Correlation Function Method(AF), etc. SA needs not only to compute derivatives of output respect to input or weights of neural network, but also to train the neural network. This will increase computational complexity. MI needs to compute density function and it is not suitable for continuous numerical values. SCM computes separation property of input-output and the correlation property of input and output attributes and is suitable for both continuous and discrete numerical values, but computation is complex. AF not only can well overcome the ID3's deficiency of tending to take value with more attributes, but also can represent the relations between all elements and their attributes. Therefore, the obtained relation degree value of attribute can reflect its importance.

AF algorithm: Suppose A is an attribute of data set D, and C is the category attribute of D. the relation degree function between A and C can be expressed as follows:

$$AF(A) = \frac{\sum_{i=1}^{n} |x_{i1} - x_{i2}|}{n} \qquad (4)$$

Where x (j = 1, 2 represents two kinds of cases) indicates that attribute A of D takes the i-th value and category attribute C takes the sample number of the j-th value, n is the number of values attribute A takes.

Then, the normalization of relation degree function value is followed. Suppose that there are m attributes and each attribute relation degree function value are

$AF(1)$, $AF(2)$, respectively. Thus, there is

$$V(k) = \frac{AF(k)}{AF(1) + AF(2) + \cdots + AF(m)} \qquad (5)$$

Which 0 < K ≤ km. Then, equation of gain can be modified as

$$Gain'(A) = (I(s_1, s_2, \cdots, s_m) - E(A)) \times V(A) \qquad (6)$$

Gain'A can be used as a new criterion for attribute selection to construct decision tree according to the procedures

of ID3 algorithm. Namely, decision tree can be constructed by selecting the attribute with the largest *Gain'A* value as test attribute. By this way, the shortcomings of using ID3 can be overcome. It construct the decision tree, this tree structure will be able to effectively overcome the inherent drawbacks of ID3 algorithm.

III. E XPERIMENTAL RESULTS

A customer database of some shopping mall is shown in Table 1 (a training sample set). The category attribute of the sample set is "buying-computer", which can take two different values: buying-computer or No buying-computer

Table 1. Shopping mall customer database

Case	Age	Color-cloth	Income	Student	Buy-computer
1	>40	Red	High	No	No
2	<30	Yellow	High	No	No
3	30--40	Blue	High	No	Yes
4	>40	Red	Medium	No	Yes
5	<30	White	Low	Yes	No
6	>40	Red	Low	Yes	No
7	30—40	Blue	Low	Yes	Yes
8	<30	Yellow	Medium	No	Yes
9	<30	Yellow	Low	Yes	No
10	>40	White	Medium	No	No

In order to illustrate the effectiveness of our present algorithm, the improved ID3 algorithm and ID3 algorithm are applied on this example to construct decision trees and comparison is made. Comparing rule extraction of the two decision tree algorithms, the resulting decision trees and classification rules by ID3 algorithm. Figure 1 and figure 2 show the generated decision trees using the ID3 algorithm and the improved ID3 algorithm, respectively.

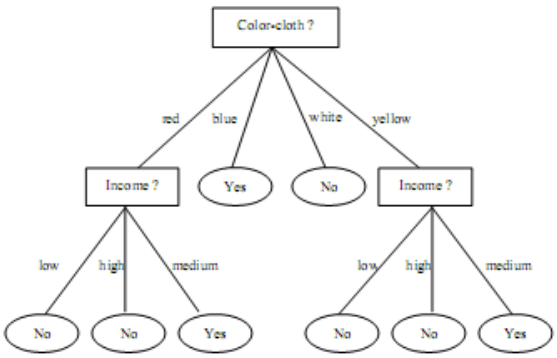


Figure 1. The obtained decision tree using ID3 algorithm

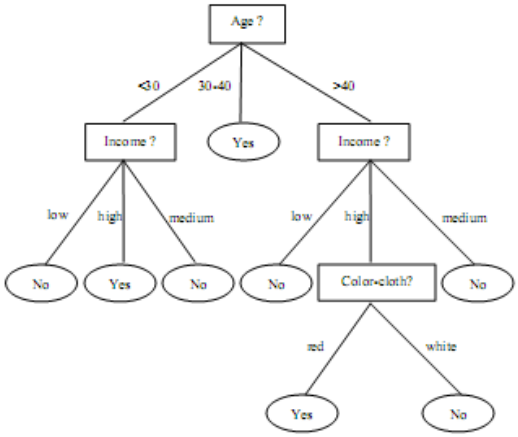


Figure 2. The obtained decision tree using improved ID3 algorithm

The two results of the experiment shows that ID3 algorithm choose attribute color-cloth as root node to generate decision tree, but the importance of attribute color-cloth is lower than the other attributes, and it is just the shortcoming of ID3 which tends to take attributes with many values. However the improved ID3 algorithm decreases the importance of attribute color-cloth in classification and comparatively enhanced the importance of attributes such as age, income, and student, etc. in classification. It well solves the problem that ID3 algorithm tends to take attributes with many values and it can obtain more reasonable and effective rules.

IV. CONCLUSION

In this project, Our Objective i.e. an improved ID3 algorithm is presented to overcome deficiency of general ID3 algorithm which tends to take attributes with many values is archived successfully. The presented algorithm makes the constructed decision tree more clear ID3 algorithm which tends to take attributes with many values. The presented algorithm makes the constructed decision tree more clear and understandable. Because it needs to compute the relation degree function value for each attribute based on ID3 algorithm, it unavoidably increases computational complexity. But with the rapid development of computer technology, the operating speed of computer gets faster and faster, the increased computational complexity can be neglected. Generally speaking, the improved ID3 algorithm takes the advantages of ID3 and AF algorithms and overcomes their disadvantages. Experiment results show that the improved ID3 can generate more optimal decision tree than general ID3 algorithm.

V.
REFERENCES

[1] I. H. Witten, E. Frank, Data Mining Practical Machine Learning Tools and Techniques, China Machine Press, 2006.

[2] Y. T. Zhang, L. Gong, Principles of Data Mining and Technology, Publishing House of Electronics Industry.

[3] D. Jiang, Information Theory and Coding [M]: Science and Technology of China University Press, 2001.

[4] S. F. Chen, Z. Q. Chen, Artificial intelligence in knowledge engineering [M]. Nanjing: Nanjing University Press, 1997.

[5] Z. Z. Shi, Senior Artificial Intelligence [M]. Beijing: Science Press, 1998.

[6] M. Zhu, Data Mining [M]. Hefei: China University of Science and Technology Press ,2002.67-72.

[7] A. P. Engelbrecht., A new pruning heuristic based on variance analysis of sensitivity information[J]. IEEE Trans on Neural Networks, 2001, 12(6): 1386-1399.

[8] N. Kwad, C. H. Choi, Input feature selection for classification problem [J],IEEE Trans on Neural Networks, 2002,13(1): 143- 159.

[9] X. J. Li, P. Wang, Rule extraction based on data dimensionality reduction using RBF neural networks [A]. ICON IP2001 Proceedings, 8th International Conference on Neural Information Processing [C]. Shanghai, China, 2001.149-153.

[10] S. L. Han, H. Zhang, H. P. Zhou, correlation function based on decision tree classification algorithm for computer application in November 2000.

IT 010

UNIVERSITY TWITTER ON CLOUD

Shantanu R. Wagh
D.Y. Patil College Of Engineering, Akurdi, Pune,
University of Pune, Maharashtra, India
Phone: +919970737151 ; E-mail: waghshantanu7@gmail.com

Hitendra A. Chaudhary
D.Y. Patil College Of Engineering, Akurdi, Pune,
University of Pune, Maharashtra, India
Phone: +919028273451 ; E-mail: hitendrachaudhary188@gmail.com

The research is sponsored by Teleparadigm Networks Pvt. Ltd. (Sponsoring information)

Abstract

In this paper, we present a project that would provide twitter like website for a university purpose intended to use by the students, professors of the colleges in the University and a University Admin. The University admin will provide with all the necessary notifications regarding the events in the University for the students and professors like result notifications, timetables, events, fees, important dates, etc. The students and staff can follow the University Admin and get the necessary notifications just by logging into their accounts and viewing their home pages. Moreover, this project would be deployed on the cloud which would make this website available 24x7 and will reduce the overhead cost of maintaining the expensive servers for the University. This project aims at utilizing the social networking and technology for educational purposes and to improve the student-teacher interaction off the classroom.

Keywords: Cloud Computing, Online Learning, Student Engagement, Twitter

1. INTRODUCTION

In existing system, there is website for a university where the updates are viewed.

The current University websites are crowded with the information concerned with all the courses and departments that University handles. Searching specific information of interest on any University website is a tedious and time consuming job. But in a typical scenario the server might be down, some information is not available on site or student community is just interested in result updates. The updates, notifications and other things are viewed by everyone. What if others are not interested in it? Staff of University might not be interested in receiving updates of a particular event and so on. For everything you need to visit University website.

In proposed system we will try to overcome some of problems encountered by creating a targeted application for a university i.e., particular to a community. For example student community can subscribe only for result updates; notifications about a particular subject and so on. We will be creating a project like twitter for University. This project will be deployed on cloud which will help in reducing the overhead cost and maintenance.

2. THEORETICAL BACKGROUND

2.1

TWITTER

Twitter is an additional way to enhance social presence. Twitter is a multiplatform Web 2.0, part social networking - part microblogging tool, freely accessibly on the Web. Other popular Web 2.0 microblogging tools include Jaiku, Tumblr, MySay, etc. Twitter, however, is one of the most popular of these microblogging tools and, therefore, was our tool of choice because it is well-established, has a large and growing participant base, interfaces well with other Web 2.0 tools, and is easily accessible.

According to the Twitter website, Twitter is a service for friends, family, and co–workers to communicate and stay connected through the exchange of quick, frequent answers to one simple question: What are you doing*?* However, the people who participate in the Twitter community—people who are geographically distributed across all continents (with North America, Europe, and Asia having the highest adoption rate)—use it for more than providing updates on their current status.

In 140 characters or less, people share ideas and resources, ask and answer questions, and collaborate on problems of practice; in a recent study, researchers found that the main communication intentions of people participating in Twitter could be categorized as daily chatter, conversations, sharing resources/URLs, and reporting news. Twitter community members post their contributions via the Twitter website, mobile phone, email, and instant messaging—making Twitter a powerful, convenient, community-controlled micro sharing environment. Depending on whom you choose to follow (i.e., communicate with) and who chooses to follow you, Twitter can be effectively used for professional and social networking because it can connect people with like interests. And all of this communication happens in real-time, so the exchange of information is immediate. (Dunlap 2009)

2.2. *CLOUD COMPUTING*

Nowadays, the term “cloud computing” has been an important term in the world of Information Technology (IT). Cloud computing is a kind of computing which is

highly scalable and use virtualized resources that can be shared by the users. Users do not need any background knowledge of the services. A user on the Internet can communicate with many servers at the same time and these servers exchange information among themselves. Cloud Computing is currently one of the new technology trends (broadband internet, fast connection and virtualization) will likely have a significant impact on teaching and learning environment. Senior people in charge of their business place challenge how to redesign their IT operations to support their business units in the light of different technology trends so they can achieve their corporate objectives. tables and figures you insert in your document are only to help you gauge the size of your paper, for the convenience of the referees, and to make it easy for you to distribute preprints.

Cloud computing is becoming an adoptable technology for many of the organizations with its dynamic scalability and usage of virtualized resources as a service through the Internet. It will likely have a significant impact on the educational environment in the future. Cloud computing is an excellent alternative for educational institutions which are especially under budget shortage in order to operate their information systems effectively without spending any more capital for the computers and network devices. Universities take advantage of available cloud-based applications offered by service providers and enable their own users/students to perform business and academic tasks. In this paper, we will review what the cloud computing infrastructure will provide in the educational arena, especially in the universities where the use of computers are more intensive and what can be done to increase the benefits of common applications for students and teachers.

3. EDUCATIONAL USAGE OF CLOUD COMPUTING

The Cloud delivers computing and storage resources to its users/customers. It works as a service on demand policy. Cloud computing is a new business model wrapped around new technologies like virtualization, SaaS and broadband internet. Recent interests offered new applications and elastic scalability with higher computing parameters. So that, these positive effects have shifted to outsourcing of not only equipment setup, but also the ongoing IT administration of the resources as well.

Refer Figure 1.

Many technologies that were previously expensive or unavailable are now becoming free to anyone with a web browser. This is true for all web sites, blogs, video sharing, music sharing, social sharing, collaboration software, editing/presentation and publishing, and computing platforms in the “cloud”. Students are already using many of these technologies in their personal lives. In the professional world, the trend of discovering and using technologies in our personal life is called “consumerization”. This means we should demand and consume the required services. Our education system should take advantage of this same trend, which will both enrich our student’s technology enabled education, and importantly, reduce the budget impact in academic institutions. University management should identify and leverage emerging technologies that are cost-effective, and strive for the broadest feasible and equitable access to technology

for students and staff. The need for hardware and software isn’t being eliminated, but it is shifting from being on-premises to being in the cloud. All that is needed

is a cheap access device and a web browser, broadband in the schools, perhaps wireless hotspots. (Ercan 2010)

3.1 Five Key Characteristics

3.1.1 Rapid elasticity: Capabilities can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down.

3.1.2 Ubiquitous network access**:** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones and laptops, etc.).

3.1.3 Pay per use**:** Capabilities are charged using a metered, fee-for-service, or advertising based billing model to promote optimization of resource use.

3.1.4 On-demand self-service**:** With many cloud computing services, a consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed without requiring human interaction with each service's provider.

3.1.5 Location independent data centre's**:** The provider's computing resources are usually pooled to serve all consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

4. TWITTER FOR EDUCATION

Faculties have recently begun experimenting with how to use Twitter in the "classroom". Communication faculties are not the only one's using Twitter in the classroom. The following describes the students' typical experiences using Twitter:

- A student is reading something in the textbook and has a question about the chapter on multimodal learning. He immediately tweets (i.e., posts) his question to the Twitter community, and gets three responses within ten minutes)—two responses from classmates, and one from her professor. This leads to several subsequent posts, including comments from two practicing professionals.
- A student is working on an assignment and is wondering about embedding music into a slideshow presentation. He tweets a question to the group and gets a response from his professor and a practicing professional. Both point the student to different online resources that explain how to embed music and provide examples to deconstruct. Within a half hour, the student has embedded music in his slideshow presentation.
- A student sends a private tweet (i.e., a private message that only the named recipient receives) to project guide regarding a difficult situation with a project team member. While in the middle of a departmental meeting, project guide immediately tweets back, arranging a time to talk with the student outside of Twitter.
- A student cannot believe what she has just heard on the news regarding federal funding of higher education and needs to share. She tweets her comment, and immediately connects with others who cannot believe it either.
- A student finds a great video about storyboarding on YouTube and posts the URL to Twitter. Her find is retweeted (i.e., reposted) three times because others also think the video is great and worth sharing.

- Joni and Patrick, who are both away at conferences, tweet various updates about what they are hearing and seeing at the conference.
- Several of the students are posting comments to Twitter while they watch a political debate. They provide commentary, along with several thousand others who are also in Twitter while watching the debate.
- A student tweets that he just posted a new entry to his blog on how vision trumps all other senses during instruction and provides the URL. His classmates, as well as other practicing professionals, read his blog post. He receives three tweets thanking him for sharing his ideas.
- As part of a research project on legacy systems, a student poses a question to the Twitter community regarding the prevalent need for COBOL programmers. She receives responses from several IT professionals, some with links to helpful resources and contacts that can help her with research.
- A student tweets that she is tired and going off to bed. She receives two tweets back from classmates wishing her a good night.

Through the use of Twitter in this way as a tool that enables just-in-time communication with the local and global (practicing professionals) community, student are able to engage in sharing, collaboration, brainstorming, problem solving, and creating within the context of their moment-to-moment experiences. Because of Twitter's ability to enable persistent presence, our social interactions occur more naturally and immediately.

5. OTHER INSTRUCTIONAL BENEFITS OF TWITTER

Besides the benefit of enhancing the potential for positive social presence during online learning opportunities, Twitter has other instructional benefits.

5.1 Addressing Student Issues in a Timely Manner

Students can use Twitter for time-sensitive matters: to ask for clarification on content or assignment requirements, notifying of personal emergencies, and alerts to issues that need some attention and action. Even though we log into the LMS several times a day, this immediate communication allows to attend the issues in a timely manner.

5.2 Writing Concisely

Because a tweet is limited to 140 characters, this encourages students to write clearly and concisely. Although a very informal writing style, it is a professionally useful skill for students to develop, especially given the growing popularity of this category of communication tool.

5.3 Writing for an Audience

Although Twitter elicits open sharing and an informal writing style, it is nevertheless critical to know your audience and share accordingly. Participating in the Twitter community helped our students learn to be sensitive to their audience, and make professional decisions about what perspectives and ideas they should publically contribute and what perspectives and ideas should remain private.

5.4 Connecting with a Professional Community of Practice

A great benefit of participating in Twitter is that many practicing professionals also participate. For example, a number of the textbook authors participate in

Twitter. Besides the networking potential, students receive immediate feedback to their questions and ideas from practicing professionals, which serves to reinforce the relevance of Twitter participation and enhance their understanding of our course content and their enculturation into the professional community of practice.

5.5 Supporting Informal Learning

Informal learning involves "activities that take place in students' self-directed and independent learning time, where the learning is taking place to support a formal program of study, but outside the formally planned and tutor-directed activities". Twitter was one tool that students used to support their informal learning activities. Through their participation in the Twitter community, they discovered resources and tools that they effectively applied to their coursework.

5.6 Maintaining On-going Relationships

Student and faculty use of Twitter is not bound by the structure of an LMS or the timing of a semester. Twitter enables faculty and students to maintain on-going relationships after a course ends. Although the semester is over, they are still in daily communication with several students from the courses. This allows continuing to advise students academically and professionally. It has also allowed for a much more natural and organic progression of relationships; instead of severing the connections at the end of the semester, we are able to continue to be in community together, learning from each other and sharing our moment-to-moment experiences.

5.7 Possible Drawbacks of Twitter

Like most, if not all Web 2.0 tools, Twitter is not appropriate for all instructional situations. For instance, Grosseck & Holotescu (2008) identify a number of problems with using Twitter for educational purposes. For instance, Twitter can be time-consuming, addictive, and possibly even encourage bad grammar as a result of its 140-character limit. Further, while Twitter is free to use on a computer connected to the Web (whether accessed via a web browser or a Twitter client like Twirl), faculty and students might be charged texting or data fees if they access Twitter on their cell phone (depending on their cell phone plans).

Despite possible drawbacks like these, the instructional benefits encourage us to continue to incorporate Twitter in our online courses (as one more tool in our toolbox), and look at other Web 2.0 tools that may help us extend the instructional power of a LMS and further enhance the social presence potential of the online learning opportunities we design and facilitate Twitter in the classroom. (Dunlap 2009)

6. MATHEMATICAL MODEL

Consider the system S,

S= {U, A, Tw, Up, FL, Sf}

U= {u1, u2, u3 ….un} No. of user.

Tw= {tw1, tw2…..twn} No. of tweets.

FL= {Fl1, Fl2….. Fln} followers and following w r t the user

Sf= {sf1, sf2…….sfn} shares the files present in user list

6.1.For sending tweets to the followers

n n+i

$\sum_{i=1}^{n} Ui$ tweets then $\sum_{i=1}^{n} Twi$

Then store in database and update to followers

Else

Sending failed.

6.2.For sharing the files

$\sum_{i=1}^{n} Ui$ shares files to $\sum_{i=1}^{n} FLi$

Then upload the file.

Otherwise

File sharing Failed.

6.3.For viewing shared files

$\sum_{i=1}^{n} Sf >= 1$

Then display file.

Else

no files to view.

7. BENEFITS OF PROPOSED SYSTEM

I. Students can get all the necessary notifications of interest very easily and at one place
II. Students can stay connected with the professors out of the classroom for necessary guidance
III. Users can share files with multiple users or a single user
IV. The application will be available 24x7 as it is deployed on cloud
V. The cost of university to maintain expensive servers will be reduced

8. COMPARISON

Refer Table 1

9. CONCLUSION

We provided evidence to suggest that students and faculty were both highly engaged in the learning process through communication and connections on Twitter and students will get updated events, results, Exam details…etc. from university. This project is used to know how to work in cloud environment as the future of IT Industry is treated as cloud computing.

10. REFERENCES

Dunlap, J.C. & Lowenthal, P.R. (2009), "Using Twitter to Enhance Social Presence" in the Journal of Information Systems Education, 20(2)

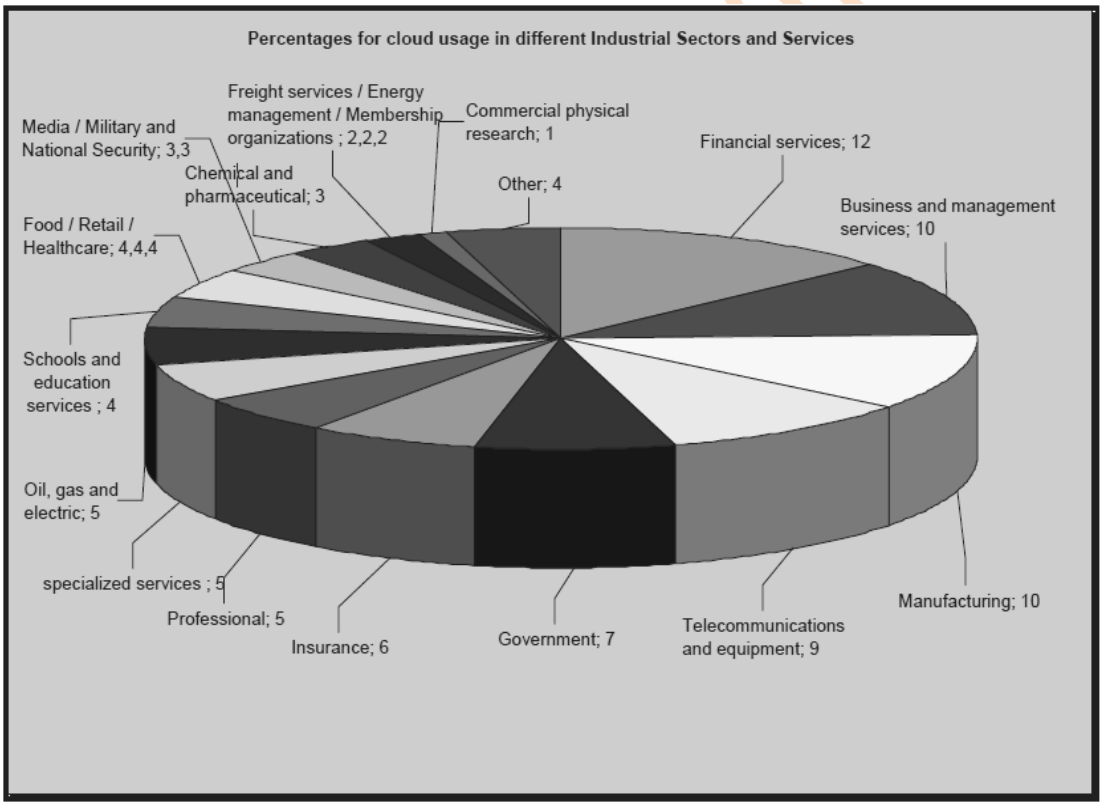
Ercan, T. (2010) "Effective Use of Cloud Computing in Educational Institutions" in Procedia Social and Behavioral Sciences 2 (2010) 938-942

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D (*et-al.*), "Above the Clouds: A Berkeley View of Cloud Computing", Electrical Engineering and Computer Sciences, University of California at Berkeley, Technical Report No. UBS/EECS-2009-28

Pocatilu, P., Alecu, F. & Vetrici, M., "Using Cloud Computing for e-Learning Systems" in Recent Advances on Data Neworks, Communications, Computers, Academy of Economic Studies, Bucharest, Romania pp.54

Grosseck, G.., Holotescu, C., "Can We Use Twitter for educational Acivities?" in the 4th International Conference eLSE "eLearning and Software for Education", Bucharest, April, 2008

Figure 1. Results of Survey in 2009



The results of a survey that have been completed in 2009 by Gartner analysts (Figure 1) about the IT trends (especially cloud computing) show that it is being used more in the areas of finance and business when compared to other sectors (Gartner, 2009). Results are shown as a pie chart and the labels on each

different slice represent different industrial sectors and services. The "/" is used to separate different sectors with the same percentage.

Table 1: Comparison between Existing and Proposed Systems

Sr no	**Existing System**	**Proposed System**
1	Intended for social networking and micro blogging	Intended for educational purpose
2	Availability problem	Available 24x7 as on cloud
3	File sharing is not allowed	File sharing is allowed
4	Maintenance cost of large servers is more	Maintenance cost of servers is minimum

IT 011

Mobile Ad-Hoc Network (MANETs)

The Art of Networking without Network

Submitted By:

Prof.Deepika A.Sarwate deepika_medhekar@yahoo.com **Contact.No : 9326972164**
JSPM NTC Rajarshri Shahu School of Computer Application Narhe Pune (MH)

Prof.Sheetal S.Patil patilsheetal@yahoo.co.in
Contact.No : 9420423491
JSPM NTC Rajarshri Shahu School of Computer Application Narhe Pune (MH)

ABSTRACT

Mobile ad hoc network (MANET) is an autonomous system of mobile nodes connected by wireless links. Each node operates not only as an end system, but also as a router to forward packets. The nodes are free to move about and organize themselves into a network. These nodes change position frequently.

To accommodate the changing topology special routing algorithms are needed. For relatively small networks flat routing protocols may be sufficient. However, in larger networks either hierarchical or geographic routing protocols are needed. There is no single protocol that fits all networks perfectly. The protocols have to be chosen according to network characteristics, such as density, size and the mobility of the nodes. MANET does not require any fixed infrastructure, such as a base station, therefore, it is an attractive option for connecting devices quickly and spontaneous.

MANETs can be used alone (for example in the military) or as a hybrid together with the Internet or other networks. Different MANET applications have different needs, and hence the various MANET routing protocols may be suitable in different areas. The size of the network and the frequency of the change in topology are factors that affect the choice of the protocols. There is no best protocol for all applications.

There is still ongoing research on mobile ad hoc networks and the research may lead to even better protocols and will probably face new challenges. The current

goal is to find an optimal balance between scalable routing and media access control, security, and service management.

INTRODUCTION

Information technology is rapidly changing from regular desktop computing, where isolated workstations communicate through shared servers in a fixed network, to an environment where a large number of different platforms communicate over multiple network platforms. In this environment the devices adapt and reconfigure themselves individually and collectively, to support the requirements of mobile workers and work teams.

In the next generation of wireless communication systems, there will be a need for the rapid deployment of independent mobile users. Mobile Ad Hoc Networks (MANETs) provide communication between all nodes in the network topology without the presence of a centralized authority; instead all nodes can function as routers. This gives the MANETs two of its most desirable characteristics; adaptable and quick to deploy. MANET research is still in progress, and currently MANETs are not widely used. Suggested areas of use will include establishing efficient communication networks for mobile workers in desolate regions or in disaster areas where existing networks have been destroyed or do not exist. To communicate in an efficient way proper routing protocols are needed.

Mobile Ad Hoc Networks (MANETs) consist of nodes that change position frequently. To accommodate the changing topology special routing algorithms are needed. For relatively small networks flat routing protocols may be sufficient. However, in larger networks either hierarchical or geographic routing protocols are needed. There is no single protocol that fits all networks perfectly. The protocols have to be chosen according to network characteristics, such as density, size and the mobility of the nodes.

Mobile ad hoc networking is one of the more innovative and challenging areas of wireless networking, one which promises to become increasingly present in our lives. Consisting of devices that are autonomously self-organizing in networks, ad hoc networks offer a large degree of freedom at a lower cost than other networking solutions.

MANETs

A MANET is an autonomous collection of mobile users that communicate over relatively “slow” wireless links. Since the nodes are mobile, the network topology may change rapidly and unpredictably over time. The network is decentralized, where all network activity, including discovering the topology

and delivering messages must be executed by the nodes themselves. Hence routing functionality will have to be incorporated into the mobile nodes.

Since the nodes communicate over wireless links, they have to contend with the effects of radio communication, such as noise, fading, and interference. In addition, the links typically have less bandwidth than a wired network. Each node in a wireless ad hoc network functions as both a host and a router, and the control of the network is distributed among the nodes. The network topology is in general dynamic, because the connectivity among the nodes may vary with time due to node departures, new node arrivals, and the possibility of having mobile nodes.

An ad hoc wireless network should be able to handle the possibility of having mobile nodes, which will most likely increase the rate at which the network topology changes. Accordingly the network has to be able to adapt quickly to changes in the network topology. This implies the use of efficient handover protocols and auto configuration of arriving nodes.

MANETs APPLICATION AREAS

Significant examples of MANETs include establishing survivable, efficient and dynamic communication for emergency/rescue operations, disaster relief efforts, and military networks. Such network scenarios cannot rely on centralized and organized connectivity, and can be conceived as applications of Mobile Ad Hoc Networks.

However, MANETs are not solely intended for disconnected autonomous operations or scaled scenarios (i.e. hundreds or even thousands of cooperation wireless nodes in a region). They may be used as hybrid infrastructure extensions and in fixed infrastructure operations. A hybrid infrastructure extension is a dynamic enhancement to a home or campus wireless networking environment. It provides extended service and allows low-cost, low complexity dynamic adjustments to provide coverage regions and range extensions away from the more fixed infrastructure backbone networks.

In hybrid infrastructure nodes move or operate on limited energy may be low-preference routing nodes, thus provides more physical stability to the overall routing grid as well. When allowing certain MANET nodes to be preferred over other nodes in a neighborhood, the more passive MANET nodes may provide range extension and dynamic routing functions on an as-needed basis. This may be appropriate within a campus, community, robotic, sensor, or localized business application. In contrast to the hybrid infrastructure extension there are no fixed access nodes or gateway points to provide confirmation coordination in a fixed infrastructure less operation. Here the participating

nodes will have to operate in a peer-to-peer fashion, with appropriate applications and protocols. Examples of esoteric ad hoc applications are ad hoc conferencing, business meeting capabilities and ad hoc homeland defense and disaster relief networks. They will require more distributed forms of auto configuration, service discovery, and management.

There are also other network application areas; cooperatives and sensors. In cooperatives a community of interest (i.e. a small town government, infrastructure-lacking world region, group of interested individuals/club) own and operate a network infrastructure together. These networks could deploy MANET technology to support a self-organizing, adaptive infrastructure. This will be desirable in disadvantaged rural regions and developing countries with lack of resources or an environment not suited for significant fixed-infrastructure developments and services. MANET technology can be used here to help building and operating inexpensive network infrastructure services. Sensor networks may be more scaled and capable using MANET technology. Commercial, environmental and military applications are all interested in this. MANET technology can support broad applications of self-organizing and distributed sensor networks.

MANET ROUTING PROTOCOLS

Generally routing protocols in MANETs are either based on the link-state (LS) routing algorithm or on the distance-vector (DV) routing-algorithm. Common for both of these algorithms is that they try to find the shortest path from the source node to the destination node. The main difference is that in LS based routing a global network topology is maintained in every node of the network. In DV based routing the nodes only maintain information of and exchange information with their adjacency nodes. Keeping track of many other nodes in a MANET may produce overhead, especially when the network is large. Therefore one of the most important issues in MANET design is to come up with schemes that will contribute to reduce routing overheads.

MANET routing protocols fall into two general categories:

- Proactive routing protocols
- Reactive routing protocols

❖ ***DIFFERENT ROUTING PROTOCOLS***

➢ **FLAT ROUTING PROTOCOLS**
- **Pro-Active / Table Driven routing Protocols**
- **Reactive / On Demand Routing Protocols**

➢ **HYBRID ROUTING PROTOCOLS**

- **HIERARCHICAL ROUTING PROTOCOLS**
- **GRAPHICAL ROUTING PROTOCOLS**

1.

FLAT ROUTING PROTOCOLS

Flat routing protocols are divided into two classes; **proactive routing (table driven) protocols** and **reactive (on-demand) routing protocols**. Common for both protocol classes is that all nodes participating in routing play an equal role. They have further been classified after their design principles; proactive routing is mostly based on LS (link-state) while on-demand routing is based on DV (distance-vector).

- **Pro-Active / Table Driven routing Protocols**

Proactive MANET protocols are table-driven and will actively determine the layout of the network. Through a regular exchange of network topology packets between the nodes of the network, a complete picture of the network is maintained at every single node. There is hence minimal delay in determining the route to be taken. This is especially important for time-critical traffic (Scientific Research Corporation, 2004).

However, a drawback to a proactive MANET of protocol is that the life span of a link is significantly short. This phenomenon is brought about by the increased mobility of the nodes, which will render the routing information in the table invalid quickly.

When the routing information becomes invalid quickly, there are many short-lived routes that are being determined and not used before they turn void. Hence, another drawback resulting from the increased mobility is the amount of traffic overhead generated when evaluating these unnecessary routes. This is especially aggravated when the network size increases. The fraction of the total control traffic that consists of actual practical data is further decreased.

Lastly, if the nodes transmit infrequently, most of the routing information is deemed redundant. The nodes, however, continue to expend energy by continually updating these unused entries in their routing tables (Scientific Research Corporation, 2004). As mentioned, energy conservation is very

important in a MANET system design. Hence, this excessive expenditure of energy is not desired.

Thus, proactive MANET protocols work best in networks that have low node mobility or where the nodes transmit data frequently.

Examples of **Proactive MANET Protocols** include:

- Optimized Link State Routing, or OLSR
- Topology Broadcast based on Reverse Path Forwarding, or TBRPF

- **Reactive / On Demand Routing Protocols**

On-demand routing is a popular routing category for wireless ad hoc routing. It is a relatively new routing philosophy that provides a scalable solution to relatively large network topologies. The design follows the idea that each node tries to reduce routing overhead by only sending routing packets when communication is requested. Common for most on-demand routing protocols are the route discovery phase where packets are flooded into the network in search of an optimal path to the destination node in the network.

There exist numerous on-demand routing protocols, but only two of them is significantly more important. These are Ad Hoc On-Demand Distance Vector Routing (AODV) and Dynamic Source Routing (DSR). These two have been chosen because both have been extensively evaluated in the MANET literature and are being considered by the Internet Engineering Task Force (IETF) MANET Working Group as the leading candidates for standardization.

➢ ***HYBRID ROUTING PROTOCOLS***

As the size of the wireless network increases, the flat routing protocols may produce too much overhead for the MANET. In this case a hierarchical solution may be preferable. CGSR, HSR, ZRP and LANMAR are four hierarchical routing protocols that have different solutions to the organization of the routing of nodes in a MANET.

-CGSR (Clusterhead-Gateway Switch Routing)
-HSR (Hierarchical State Routing)
-ZRP (Zone Routing Protocol)

-LANMAR (Landmark Ad Hoc Routing Protocol)

➢ ***GEOGRAPHICAL ROUTING PROTOCOLS***

There are two approaches to geographic mobile ad hoc networks:

1. Actual geographic coordinates (as obtained through GPS – the Global Positioning System).

2. Reference points in some fixed coordinate system.

An advantage of geographic routing protocols is that they prevent network-wide searches for destinations. Control and data packets can be sent in the general direction of the destination if the recent geographical coordinates are known. This reduces control overhead in the network. A disadvantage, however, is that all nodes must have access to their geographical coordinates all the time to make the geographical routing protocols useful. The routing update must be done faster than the network mobility rate to make the location-based routing effective. This is because the nodes' locations may change quickly in a MANET.

GeoCast (Geographic Addressing and Routing)
DREAM (Distance Routing Effect Algorithm for Mobility)
GPSR (Greedy Perimeter Stateless Routing)

CONLUSION

All the routing protocols mentioned in this essay are either on-demand or proactive. There is a trade-off between sending updates often or just when needed. Sending updates may produce overhead in mobile ad hoc networks because the nodes are moving frequently. When the size of the network is small a flat routing protocol will be sufficient. Then each node keeps track of the other nodes in its routing table. How the nodes discover other nodes and how they send requests for a destination, differs between the routing protocols.

Flat routing protocols are available immediately and they support quality of service in MANETs. However, when the size of a MANET increases the flat routing protocols may not be sufficient. Then either a hierarchical or a geographic routing protocol would be a better solution. The hierarchical routing protocols organize the nodes in hierarchies and have smaller routing tables because the nodes only need to keep track of their levels in the hierarchy. Also, in search for destinations the amount of flooding packets is reduced.

However, the hierarchical routing protocols may also produce overhead to maintain the hierarchical structure. The geographic routing protocols use the knowledge of node locations to organize the structure of a MANET. They may produce overhead when exchanging coordinates, but all in all they can become more scalable and effective than the flat routing protocols.

MANETs can be used alone (for example in the military) or as a hybrid together with the Internet or other networks. Different MANET applications have different needs, and hence the various MANET routing protocols may be suitable in different areas. The size of the network and the frequency of the change in topology are factors that affect the choice of the protocols. There is no best protocol for all applications. For flat, hierarchical and geographic routing protocols, scalability is a big challenge. There is still ongoing research on mobile ad hoc networks and the research may lead to even better protocols and will probably face new challenges. The current goal is to find an optimal balance between scalable routing and media access control, security, and service management.

REFERENCE

➢ **WEBSITES:**

- www.google.com
- www.wikipedia.com

➢ **BOOKS:**

- **Internet Protocols and Networking :: By Williams Schilling**
- **TCP/IP Protocol suit :: By Forouzon**

IT 012

MOBILE PAYMENT MARKET AND RESEARCH IN INDIA: PAST, PRESENT AND FUTURE

Urvashi Kumari HOD MCA Department

Dr.D.Y Patil Institute of Management, Ambi, Pune, M.H, India

Email id: urvashijj@gmail.com

Ph no. : 9767102279

Abstract:

Like many developing countries, today India is exponentially growing in terms of mobile users which were as per Telecom Regulatory Authority of India (TRAI), about 791.2 million as at the end of February 2011 and growing at about 8 million a month. Even people in rural India are using mobile phones so mobile banking could become the currency for Indian large population. There are many initiatives taken by banking industry for the use of mobile phones for banking application. As this new concept of mobile commerce will be useful for masses, it is worthwhile to examine various components of its interface with payment system, technology, security, functionality and regulatory aspects. At this point we take a look on the current state of the mobile commerce, review literature for mobile payment services, analyze various factors that affect the market for mobile commerce and give future direction of research for still emerging field. Researcher found that while awareness for mobile payment in India started in 2000 and never took off, still the time is not matured for m-payment.

Keywords: Mobile payment systems, Mobile payment research, Mobile commerce, TRAI

1. Introduction

During last 20 years, mobile telephony has changed the way people communicate and work with mobile phones which are fully equipped with functionalities. These functionalities of mobile telephones are much more than needed of a simple telephone, which have motivated many service providers for value added

services, use of mobile phones to store and access information and mobile commerce in general. With the widespread use of mobile devices, a new type of channel, called mobile commerce, is emerging which is a successor of E-Commerce. People from every age group are using mobile phones not only for telephony but for entertainment, internet access, mails, mobile learning and many more applications. The behavior of mobile user opens profitable opportunities to service providers and business group. Mobile phones which are used more than any other device by the masses even in rural India can be used to market, sell, produce and provide product and services. Mobile payments services have plentiful advantages over traditional payment methods. A mobile payment or m-payment may be defined, for our purposes, as any payment where a mobile device is used to initiate, authorize and confirm an exchange of financial value in return for goods and services (Au and Kauffman, 2007). Apart from their evident flexibility, they allow consumers who do not have easy access to banking facilities to participate readily in financial transactions. Unfortunately, offered mobile payment solutions in India are not interoperable; i.e. they only put forward services for merchants registered with them and do not allow the transfer of money to, or between, users of other payment providers. This limitation reduces the widespread adoption of mobile payments. Initially fixed line telephone billing system was used to charge mobile telephony for the products and services and the deployment of mobile telecom billing system is still very classic way to charge for mobile commerce transaction. However, telephony billing system has its own limitation of high payment transaction fees, product and service provider complaints about unjust profit sharing and requirement of billing services with limited roaming of mobile networks. Mobile payments also attracted researchers, e.g., Dahlberg et al. (2003a; 2003b), Ondrus and Pigneur (2004), Pousttchi (2003), and Zmijewska et al. (2004b). Hundreds of mobile payment services as well as access to electronic payment and Internet banking were introduced all over the world .Ondrus J. & Pigneur Y. (2006) states that Mobile payments have the potential to revolutionize methods of paying products and services.. The scope of this paper is restricted only to Indian markets where only banks are authorized to offer mobile payment services.

2. Literature Review

Mobile devices can be used in a variety of payment scenarios such as payment for digital content (e.g. ring tones, logos, news, music, or games), concert or flight tickets, parking fees, and bus, tram, train and taxi fares, or to access and use electronic payment services to pay bills and invoices. Payments for physical goods are also possible, both at vending and ticketing machines, and at manned Point-of-Sale terminals. Typical usage entails the user electing to make a mobile

payment, being connected to a server via the mobile device to perform authentication and authorization, and subsequently being presented with confirmation of the completed transaction (Antovski & Gusev, 2003; Ding & Hampe, 2003). Several mobile payment companies and initiatives in Europe have failed and many have been discontinued (Dahlberg et al., 2007). In Europe and North America with few exceptions such as Austria and Spain the development of mobile payments has not been successful. However, mobile payment services in Asia have been fairly successful especially in South Korea, Japan and other Asian countries (e.g., Mobile Suica, Edy, Moneta, Octopus, GCash). Mobile Payment, worldwide has failed and as Booz Allen and Hamilton points out, the only way to make mobile payment work is increase cooperation among the key stakeholders – i.e. payment industry (banks/credit card issuers), operators, handset manufacturers, merchants, channel enablers and of course the regulatory authorities. NTT DoCoMo has 20 million subscribers and 1.5 million of them have activated credit card functionality in Japan. There are 100,000 readers installed in Japan (Ondrus and Pigneur, 2007). The main difference between successful implementations of mobile payment services in the Asia Pacific region and failure in Europe and North America is primarily attributed to the 'payment culture' of the consumers that are country-specific.

3. Mobile Payment Solutions

Mobile payment solutions may be classified according to the type of financial rules and regulations followed in a country. There are three types of mobile payment markets (Carr, M, 2009)

(1) Bank account based: This is for highly regulated markets, where only the banks are entitled to offer mobile payment services. Banks have several million customers and telecommunication operators also have several million customers. If they both collaborate to provide an m-payment solution it is a win-win situation for both industries. In this model, the bank account is linked to the mobile phone number of the customer. When the customer makes an m-payment transaction with a merchant, the bank account of the customer is debited and the value is credited to the merchant account. In India, when the customer makes a payment from the mobile, the bank account of the customer is debited

(2) Credit card Based: In the credit card based m-payment model, the credit card number is linked to the mobile phone number of the customer. When the customer makes an m-payment transaction with a merchant, the credit card is

charged and the value is credited to the merchant account. Credit card based solutions have the limitation that it is heavily dependent on the level of penetration of credit cards in the country. In India, the number of credit card holders is 15 million (Subramani, 2006). Only this small segment of the population will benefit in the credit card based model. Though limited in scope, there may be high demand within this segment for a payment solution with credit cards and also, may provide high volumes of transactions.

(3) Telecommunication company billing based: Minimally regulated markets, which allow Telecommunication Service Providers (TSPs) to handle the subscribers' cash with a TSP account, and allow the TSP to accept and disburse cash from its outlets. Customers may make payment to merchants using his or her mobile phone and this may be charged to the mobile phone bills of the customer. The customer then settles the bill with the telecommunication company (Zheng and Chen, 2003). This may be further classified into prepaid airtime (debit) and postpaid subscription (credit).

4. Mobile Payment Technologies

The mobile technology background provides various possibilities for implementing m-payments. Fundamentally, a GSM mobile phone may send or receive information (mobile data service) through three possible channels – SMS, USSD or WAP/GPRS. The choice of the channel influences the way m-payment schemes are implemented. Secondly, the m-payment client application may reside on the phone or else it may reside in the subscriber identity module (SIM). We briefly describe NFC technology as another possibility.

Short Message Service (SMS)

This is a short text message service that allows short messages (140-160 characters) to be sent from a mobile phone or from the web to another mobile phone. Short messages are stored and forwarded by SMS centers. SMS messages have a channel of access to phone different from the voice channel (Valcourt, Robert and Beaulieu, 2005). SMS can be used to provide information about the status of one's account with the bank (informational) or can be used to transmit payment instructions from the phone (transactional) by paying ordinary SMS charges.

For example State Bank of India (SBI) is offering following functionalities with SMS :

- Enquiry Services (Balance Enquiry/Mini Statement)
- Mobile Top up
- DTH Top up/ recharge
- IMPS- Mobile to Mobile Transfer
- Change MPIN

Business Rules

- All Current/ Savings Bank Account holders in P segment are eligible.
- Transaction limit per customer per day is Rs.1,000/- with a calendar month limit of Rs.5,000/-
- All customers can avail the Service irrespective of telecom service provider.
- The Service is free of charge. SMS cost will be borne by the customer.
- As a matter of abundant precaution, Customers are requested to delete all the messages sent to the number 9223440000, once the response for their request has been received.

Unstructured Supplementary Services Delivery (USSD)

Unstructured Supplementary Service Data (USSD) is a technology unique to GSM. It is a capability built into the GSM standard for support of transmitting information over the signaling channels of the GSM network. USSD provides session-based communication, enabling a variety of applications. USSD is session oriented transaction-oriented technology while SMS is a store-and-forward technology. Turnaround response times for interactive applications are shorter for USSD than SMS.

For instance SBI is offering following functionalities to M payment users:

- Enquiry Services (Balance Enquiry/Mini Statement)
- Mobile Top up
- Funds Transfer (within Bank)

Business Rules

- All Current/ Savings Bank Account holders in P segment are eligible.
- Transaction limit per customer per day is Rs.1,000/- with a calendar month limit of Rs.5,000/-
- The Service is available for subscribers of select telecom operators only.
- The Service is free of charge. USSD session charges will be borne by the customer.

- The service is session based and requires a response from the user within a reasonable time.

WAP/GPRS

General Packet Radio Service (GPRS) is a mobile data service available to GSM users. GPRS provides packet-switched data for GSM networks. GPRS enables services such as Wireless Application Protocol (WAP) access, Multimedia Messaging Service (MMS), and for Internet communication services such as email and World Wide Web access in mobile phones. Billing through WAP is used by the consumers to buy content and services like ring tones, mobile games and wall papers from WAP sites that is charged directly to their mobile phone bill. It is an alternative payment mechanism for SMS billing.

Phone-based Application (J2ME/BREW)

The client m-payment application can reside on the mobile phone of the customer. This application can be developed in Java (J2ME) for GSM mobile phones and in Binary Runtime Environment for Wireless (BREW) for CDMA mobile phones. Personalization of the phones can be done over the air (OTA).

SIM-based Application

The subscriber identity module (SIM) used in GSM mobile phones is a smart card i.e., it is a small chip with processing power (intelligence) and memory. The information in the SIM can be protected using cryptographic algorithms and keys. This makes SIM applications relatively more secure than client applications that reside on the mobile phone. Also, whenever the customer acquires a new handset only the SIM card needs to be moved (Card Technology, 2007). If the application is placed on the phone, a new handset has to be personalized again.

Near Field Communication (NFC)

NFC is the fusion of contactless smartcard (RFID) and a mobile phone. The mobile phone can be used as a contactless card. NFC enabled phones can act as RFID tags or readers. This creates opportunity to make innovative applications especially in ticketing and couponing (Ondrus and Pigneur, 2007). The 'Pay-Buy Mobile' project launched by the GSM Association (fourteen mobile operators are part of the initiative) targets 900 million mobile users with a common global approach using NFC (Card Technology Today, 2007).

Dual Chip

Usually the m-payment application is integrated into the SIM card. Normally, SIM cards are purchased in bulk by telecom companies and then customized for use before sale. If the m-payment application service provider has to write an m-payment application in the SIM card, this has to be done in collaboration with the telecommunications operator (the owner of the SIM). To avoid this, dual chip phones have two slots one for a SIM card (telephony) and another for a payment chip card. Financial institutions prefer this approach as they can exercise full control over the chip and the mobile payment process (Karnouskos and Fokus, 2004). But, customers would have to invest in dual chip mobile devices.

Mobile Wallet

A m-payment application software that resides on the mobile phone with details of the customer (and his or her bank account details or credit card information) which allows the customer to make payments using the mobile phone is called as a mobile wallet. Customers can multi-home with several debit or credit payment instruments in a single wallet. Several implementations of wallets that are company-specific are in use globally.

5. Conclusion:

While awareness for mobile payment in India started in 2000 and never took off, still the time is not matured for m-payment. But India is hotbed for mobile payment services because of following reasons:

- While PC users in India is close to 180 million+ , the mobile population stands at **791.2 milllion+ [adds 8 million subscribers every** month**].**
- Mobiles are available at as low as Rs 1200-1500 [i.e. 30-35$]
- Expected growth of mobile users expected to reach 900+ million in the next 2 years

With India being a hot market, currently there are SMS based services(Paymate / mChek) and GPRS based services (JiGrahak) is available . While SMS based M payment works on any phone with SMS capability, GPRS based mobile phones will need Java-enabled phone and software is also required to download. The early birds of m payment will be the users who are most comfortable with online payments, technically aware of mobile payment, well-to-do guys who will carry the high end phones and will look for secure payment solution.

6. Future research in M-payment:

The unanswered questions about M-payment which needs to be addressed for further research are as follows:

The paying behavior of the consumer in India is still by cash or cheque. Even credit card has not got its popularity then why do people switch to m payment mode? What is the motivation for the merchants and the service providers like Paymate and JiGrahak for generating revenue as users are using these without any charges. Moreover, the situation today in India is not very clear whether or not mobile payments are on their way to becoming a standard payment service. Without conducting a structured field analysis involving practitioners of different industries active in mobile payments, it is very difficult to get a good picture of the reality

7. References:

A.S. Lim (2007). Inter-consortia battles in mobile payments standardisation, Electronic Commerce Research and Applications (2007), doi:10.1016/j.elerap.2007.05.003

Antovski, L., & Gusev, M. (2003). M-Payments. Paper presented at the 25th International

Conference of Information Technology Interfaces, Cavtat, Croatia

Carr, M., Framework for Mobile Payment Systems in India, Advanced EBusiness Methods. Edited By: Milena M. Head Eldon Y. Li, Information Science Reference, 2009.

Dahlberg, T., Mallat, N., & Öörni, A. (2003a). Consumer acceptance of mobile payment

solutions - ease of use, usefulness and trust. Paper presented at the 2nd International

Conference on Mobile Business, Vienna, Austria, June 23-24.

Dahlberg, T., Mallat, N., & Öörni, A. (2003b). Trust enhanced technology acceptance model

- consumer acceptance of mobile payment solutions. Paper presented at the 2nd Mobility

Roundtable, Stockholm, Sweden, May 22-23

Deepti Kumar_, Timothy A Gonsalves†, Ashok Jhunjhunwala‡ and Gaurav Raina§

Mobile Payment Architectures for India

Ding, M. S., & Hampe, J. F. (2003). Reconsidering the Challenges of mPayments: A

Roadmap to Plotting the Potential of the Future mCommerce Market. Paper presented at

the 16th Bled eCommerce Conference, Bled, Slovenia, June 9-11.

E. Valcourt, J. Robert, & F. Beaulieu, (2005). Investigating mobile payment: supporting technologies, methods, and use. IEEE International Conference on Wireless And Mobile Computing, Networking And Communications, (WiMob'2005), Aug. 2005 Page(s):29 - 36 Vol. 4 Digital Object Identifier 10.1109/WIMOB.2005.1512946

Ondrus, J., & Pigneur, Y. (2004). Coupling Mobile Payments and CRM in the Retail

Industry. Paper presented at the IADIS International Conference e-Commerce, Lisbon,

Portugal, Dec. 14-16.

Ondrus J. & Pigneur Y. (2006). Towards A Holistic Analysis of Mobile Payments: A Multiple Perspectives Approach. Journal Electronic Commerce Research and Applications, 5(3), 246-257.

X. Zheng & D.Chen (2003). Study of mobile payments systems. IEEE International Conference on E-Commerce, CEC 2003, June 2003 Page(s):24 – 27 Digital Object Identifier 10.1109/COEC.2003.1210227

Y.A. Au & R.J. Kauffman, (2007). The economics of mobile payments: Understanding stakeholder issues for an emerging financial technology application, Electronic Commerce Research and Applications, doi:10.1016/j.elerap.2006.12.004

IT 013

STEGANOGRAPHY USING AUDIO VIDEO FILES

Full name: Sayli S. Kankam.,Sayali Kshirsagar, Sneha Sinha, Chaitali Patil

Designation: Student.

Name of the organization: College of engineering Manjari, Pune.

City: Pune.

State: Maharashtra.

Country: India.

E-mail id: sayli_k@ymail.com.

Mobile/landline phone number: 9892160675.

ABSTRACT

Computer technology and the Internet have made a breakthrough in the existence of data communication. This has opened a whole new way of implementing steganography to ensure secure data transfer. Steganography is the fine art of hiding the information. Hiding the message in the carrier file enables the deniability of the existence of any message at all. In this era of software revolution, inspection plays a significant role. All the software development firms follow their own stringent policies to ensure software reliability. ***Approach:*** *In this paper we have presented the steganography technique for hiding the variable sized secret messages into video file. We have used encryption and compression techniques. The compression is used when the secret file is large before hiding it using LSB algorithm. The password is used as secret for encryption and decryption purpose.* ***Results:*** *We can hide both text and image file different sizes into video cover file. We have used authentication for login and logout the system for making system more secure and robust. The only authorized user can hide and disclose the message. The text and image file of different sizes are used to test the system. We found that the system satisfy all requirements of steganography .The system is secured and more robust.* ***Conclusion:*** *In this study, we have concluded that the LSB algorithm is the most efficient and common approach for embedding the data. This data can be stored in either audio/video file. The proposed system provides a robust and secure way of data transmission.*

Keywords: audio/video file, encryption, decryption, LSB algorithm, steganography.

1. INTRODUCTION

Steganography is an ancient art of conveying messages in a secret way that only the receiver knows the existence of message. The subject of steganography has been brought into the limelight by several intelligence agencies and the news media in recent times. Apart from using state of the art, communication technologies and media, the agencies are using cryptography as well as steganography to aid themselves with their objectives [1]. So, a fundamental requirement for a steganographic method is imperceptibility; this means that the embedded messages should not be discernible to the human eye.

The word steganography derives from the Greek word steganos, which means covered or secret, and graphy which means writing or drawing. Steganography is also referred to as Stego. The concept of steganography has existed for thousands of years. The Greek used to pass secret information by writing in wax-covered tablets: wax was first scraped off a tablet, the secret message was written on the tablet, and then the tablet was covered again with the wax [2]. Another technique was to shave a messenger's head, tattoo a message or image on the bald head, and let hair grow again so that the tattoo could not be seen. Shaving the head again revealed the tattoo [2]. The use of invisible ink was also used extensively during the World War II. The invisible ink method and other traditional stego methods were extensively used but the invisible secret message gets revealed when heated. Then the image files are used to hide messages. But image files are not the only carriers [7]. Secret information can be hidden in computer image files (JPEG, GIF, BMP), audio files (WAV, MP3) [5], video files (MPEG, AVI), or even text files. Provided the steganographic algorithm is good enough and a Stego'd video along with the original video, even an adept steganography expert would be unable to detect the hidden information from the image. Making use of the Internet, secret information hidden in the carrier can be transmitted quickly, secretly, and securely.

Over the past few years, numerous Steganography techniques that embed hidden messages in multimedia objects have been proposed. This is largely due to the fact that multimedia objects often have a highly redundant representation which usually permits the addition of significantly large amounts of stego-data by means of simple and subtle modifications that preserve the perceptual content of the underlying cover object [7]. Hence they have been found to be perfect candidates for use as cover messages.

A message, either encrypted or unencrypted, can be hidden in a computer video file (containing the picture of, for instance, an innocent 2 year old baby) and transmitted over the Internet, a CD or DVD, or any other medium [8]. The image file, on receipt, can be used to extract the hidden message. This design incorporates the most powerful modified LSB algorithm to encode the message into video file.

Steganography Vs Cryptography -Steganography is not an alternative to cryptography [1]. Steganography is the dark cousin of cryptography. While cryptography provides privacy, steganography is intended to provide secrecy. In other words, cryptography works to mask the content of a message; steganography works to mask the very existence of the message.

REQUIREMENTS OF STEGANOGRAPHY SYSTEM

There are many different protocols and embedding techniques that enable us to hide data in a given object. However, all of the protocols and techniques must satisfy a number of requirements so that steganography can be applied correctly .The following is a list of requirements that steganography techniques must satisfy.

The integrity of hidden information after it has been embedded inside the Stego object must be correct. The secret message must not be changed in any way such as additional information being added, loss of information or changes to the information after it has been hidden. If secret information is changed during steganography, it would defeat the whole point of process.
The stego object must remain unchanged or almost unchanged to the naked eye. If the stego object changes significantly and can be noticed, a third party may see that information is being hidden and therefore could attempt to extract or destroy it.
In steganography. Changes in the stego object must have no effect on the message. Imagine if you had an illegal copy of an image that you would like to manipulate in various ways. These manipulations can be simple processes such as resizing, trimming or rotating the image. The Stegano inside the image must survive these manipulations.
Otherwise the attackers can be very easily removing the Stegano and point of steganography will be broken.
Finally, we always assume that the attacker knows that there is hidden information inside the steno object.

2. = STEGANOGRAPHIC TECHNIQUES

Over the past few years, numerous steganography techniques that embed hidden messages in multimedia objects have been proposed. There have been many techniques for hiding information or messages in images in such a manner that the alterations made to the image are perceptually indiscernible. Common approaches are include [10]:

(i) Least significant bit insertion (LSB)

(ii) Masking and filtering

(iii) Transform techniques

Least significant bits (LSB) insertion is a simple approach to embedding information in image file. The simplest steganographic techniques embed the bits of the message directly into least significant bit plane of the cover-image in a deterministic sequence. Modulating the least significant bit does not result in human perceptible difference because the amplitude of the change is small.

Masking and filtering techniques, usually restricted to 24 bits and gray scale images, hide information by marking an image, in a manner similar to paper watermarks. The techniques performs analysis of the image, thus embed the information in significant areas so that the hidden message is more integral to the cover image than just hiding it in the noise level.

Transform techniques embed the message by modulating coefficients in a transform domain, such as the Discrete Cosine Transform (DCT) used in JPEG compression, Discrete Fourier Transform, or Wavelet Transform. These methods hide messages in significant areas of the cover-image, which make them more robust to attack. Transformations can be applied over the entire image, to block throughout the image, or other variants.

3. PROPOSED SYSTEM

There are lots of steganographic programs available. A few of them are excellent in every respect; unfortunately, most of them lack usable interfaces, or contain too many bugs, or unavailability of a program for other operating systems. The proposed application will take into account these shortcomings, and since it will be written in Java, operability over multiple operating systems and even over different hardware platforms would not be an issue. This proposed stego machine provides easy way of implementing the methods. The idea behind this design is to provide a good, efficient method for hiding the data from hackers

and sent to the destination securely. This system would be mainly concerned with the algorithm ensuring the secure data transfer between the source and destination. This proposed system is based on video steganography for hiding data in the video image, retrieving the hidden data from the video using LSB (Least Significant Bit) modification method. This design looks at a specific class of widely used image based steganographic techniques, namely LSB steganography and investigate under what conditions can an observer distinguish between stegoimages (images which carry a secret message) and cover images (images that do not carry a secret message).

Carrier Image Secret Message

S

tego'd Image

Fig. 1 Steganography using video image

Steganography Terms:

Cover-Medium – The medium in which information is to be hidden, also sometimes called as Cover image or carrier.

Stego-Medium – A medium in which information is hidden

Message – The data to be hidden or extracted

In summary:

Stego_medium = hidden_message + carrier + stego_key

A. Least Significant Bit (LSB) Modification Method

The least significant bit (LSB) algorithm is used in this stego machine to conceal the data in a video file. The main advantage of the LSB coding method is a very high watermark channel bit rate and a low computational complexity. The

robustness of the watermark embedded using the LSB coding method, increases with increase of the LSB depth is used for data hiding. In this method, modifications are made to the least significant bits of the carrier file's individual pixels, thereby encoding hidden data [6]. Here each pixel has room for 3 bits of secret information, one in each RGB values. Using a 24-bit image, it is possible to hide three bits of data in each pixel's color value using a 1024x768 pixel image; also it is possible to hide up to 2,359,296 bits. The human eye cannot easily distinguish 21-bit color from 24-bit color [3]. As a simple example of LSB substitution, imagine "hiding" the character 'A' across the following eight bytes of a carrier file:

(00100111 11101001 11001000)

(00100111 11001000 11101001)

(11001000 00100111 11101001)

Letter 'A' is represented in ASCII format as the binary string 10000011.

These eight bits can be "written" to the LSB of each of the eight carrier bytes as follows (the LSBs are italicized and bolded):

(0010011***1*** 1110100***0*** 1100100***0***)

(0010011***0*** 1100100***0*** 1110100***0***)

(1100100***1*** 0010011***1*** 11101001).

With such a small variation in the colors of the video image it would be very difficult for the human eye to discern the difference thus providing high robustness to the system [4].

B.Advantages of Proposed System

The advantages of the proposed stego machine are

A very usable and good looking wizard based GUI(Graphical User Interface) for the system.
Ability to operate the system with no prior training and consultation of any help files.
Ability to conceal and reveal the exact hidden data from video file without disturbing the running application or new application.
Ability to encrypt and decrypt the data with the Images.
With this system, an image, after hiding the data, will not degrade in quality.

MODULES OF STEGO MACHINE

The video stegomachine performs the process of conceal and reveal in following modules. The modules of Video Stegomachine are:

Video Header Information
File Handling
Encryption
Steganography – Conceal data
DeSteganography - Reveal original data
Decryption
Graphical User Interface

STEGANOGRAPHY – CONCEAL DATA

Here the carrier file (AVI file) length is obtained and checked for whether it is eight times greater than that of the text file. Find the starting point of the data in the AVI file and create a key file by writing the content of the AVI file starting from the data to the end. The carrier file is converted into binary. The result is overwritten to the data part of the AVI file and as well as written into the newly created text file. The output obtained for this system is a stego'd video file, and a key file which is to be shared by a secure channel. Fig. 2 depicts the clear picture of concealing the data.

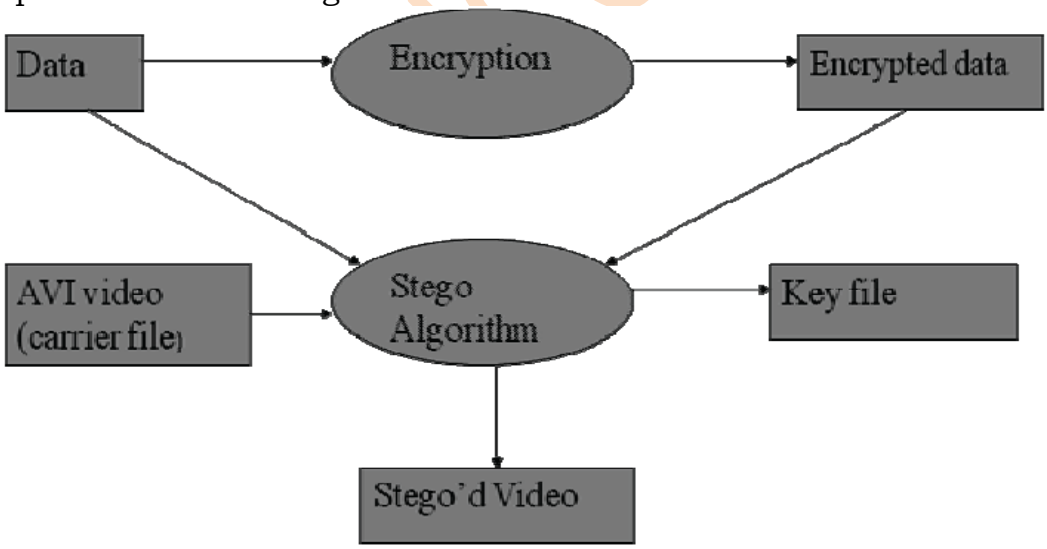


Fig.2 Steganography flowchart

DESTEGANOGRAPHY- REVEAL ORIGINAL DATA

This DeSteganography module decodes the video file to retrieve the hidden data from video. Here the carrier file (AVI file) and the Key file are given as input. The AVI file and the Key file are opened in a Random Access Mode to find the starting point of the data in the AVI file. This reads the AVI file and Key file Byte by Byte and finds the difference between them. The output obtained is an original AVI video

file, and a data file that is the message which is hidden inside the AVI video file. Fig.3 illustrates the process of revealing the original data from Stego'd video file.

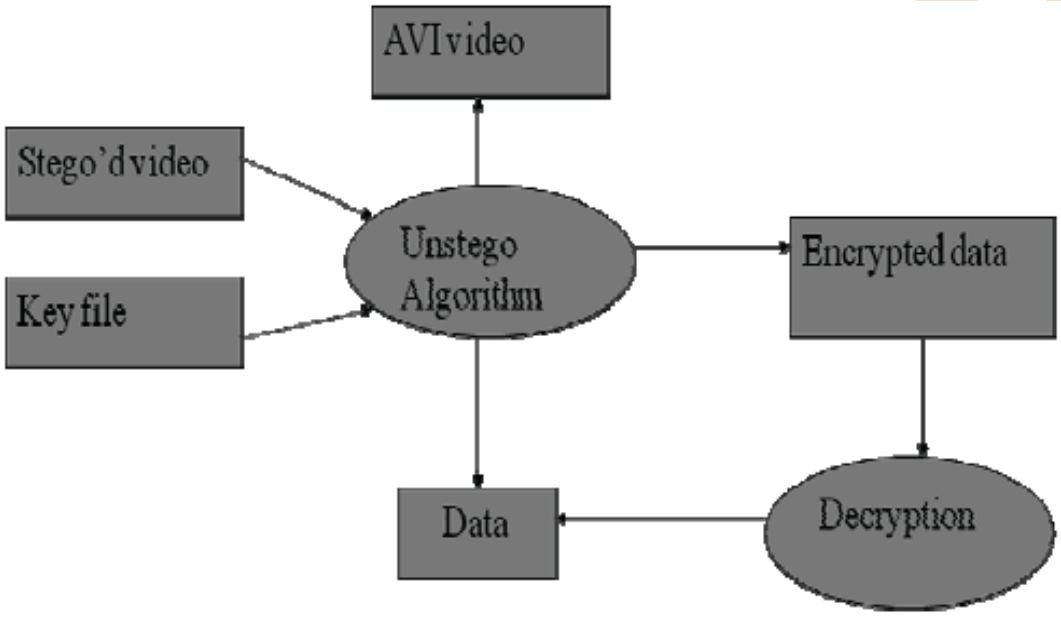


Fig.3 DeSteganography

STEGANOGRAPHY MODEL

Consider the following model for data embedding in digital images. Let {*x*1, *x*2,……, *xn*} denote *n* features (*symbols*) of the original cover image. These features could be in the spatial domain, frequency domain or a combination thereof. For example, for least significant bit encoding, *xi*'s correspond to the image pixel intensity values and for spread spectrum steganography [3] these are the discrete cosine transform coefficients. By embedding a subset of the message (perhaps modulating a message carrier) in *xi* let the resulting distortion induced to the host feature be *di*. Note that the total embedding induced distortion is

usually bounded above, say, by *D*, due to perceptual and information theoretic considerations. Let the number of message bits embedded in *ith* original image symbol be *mi*. We can formulate the *steganography problem* using 0-1 decision

variables *yi* where

yi = { 1 if symbol *xi* is selected for embedding

0 otherwise

Then, the steganography problem can be stated as follows:

maximize i=1to n$\Sigma miyi$ ..... (1)

subject to

i=1to n$\Sigma diyi \leq D$.....(2)

$yi \in \{0, 1\}; i = 1, 2,.....,n.....$ (3)

The goal is to maximize the length of the embedded message size in the image, given by the objective function (1). Constraint (2) ensures that the embedding induced distortion does not exceed the specified bound *D*. Here, we have implicitly assumed that each host symbol can carry a variable number of message size. Without loss of generality we also make the following assumptions about the coefficients *mi*, *di* and *D*.

These coefficients are non-negative integers; fractional values can be transformed to integers after multiplication by a suitable factor.
$mi > 0$, for all i, if not, the corresponding feature need not be considered.
$0 < di < D$. If $di = 0$ then it does not affect the solution and, if $di > D$ the corresponding image source symbol *xi* will not be considered as a candidate for embedding.
$0 < D <$ i=1 to nΣ *di*. For, if $D = 0$ then we get the trivial solution by setting $yi = 0$, for all i and, if $D \geq i$=1 to nΣ *di* set $yi = 1$, for all i.

9. TECHNICAL SPECIFICATION

A. Advantage:

Concealing messages within the lowest bits of noisy images or sound files.

Concealed messages in tampered executable files, exploiting redundancy in the targeted instruction set.

Pictures embedded in video material (optionally played at slower or faster speed).

Secure Steganography for Audio Signals.

Steganography system plays an important role for security purpose

B. Disadvantages:

It provides the storing of data in an unprotected mode.

Password leakage may occur and it leads to the unauthorized access of data.

C. Applications:

Confidential communication and secret data storing

The "secrecy" of the embedded data is essential in this area. Historically, steganography have been approached in this area. Steganography provides us with:

Potential capability to hide the existence of confidential data

Hardness of detecting the hidden (i.e., embedded) data

Strengthening of the secrecy of the encrypted data

In practice, when you use some steganography, you must first select a vessel data according to the size of the embedding data. The vessel should be innocuous. Then, you embed the confidential data by using an embedding program (which is one component of the steganography software) together with some key. When extracting, you (or your party) use an extracting program (another component) to recover the embedded data by the same key (“common key" in terms of cryptography). In this case you need a "key negotiation" before you start communication.

Attaching a stego file to an e-mail message is the simplest example in this application area. But you and your party must do a "sending-and-receiving"

action that could be noticed by a third party. So, e-mailing is not a completely secret communication method.

There is some other communication method that uses the Internet Webpage. In this method you don't need to send anything to your party, and no one can detect your communication. Each secrecy based application needs an embedding process which leaves the smallest embedding evidence. You may follow the following.

(A) Choose a large vessel, larger the better, compared with the embedding data.

(B) Discard the original vessel after embedding.

For example, in the case of Qtech Hide & View, it leaves some latent embedding evidence even if the vessel has a very large embedding capacity. You are recommended to embed only 25% or less (for PNG / BMP output) of the maximum capacity, or only 3% of the vessel size (for JPEG output).

ii. Protection of data alteration

We take advantage of the fragility of the embedded data in this application area.“The embedded data can rather be fragile than be very robust." Actually, embedded data are fragile in most steganography programs. Especially, Qtech Hide & View program embeds data in an extremely fragile manner. However, this fragility opens a new direction toward an information-alteration protective system such as a "Digital Certificate Document System." The most novel point among others is that "no authentication bureau is needed." If it is implemented, people can send their "digital certificate data" to any place in the world through Internet. No one can forge, alter, nor tamper such certificate data. If forged, altered, or tampered, it is easily detected by the extraction program.

FUTURE SCOPE

At present we are hiding the data in compressed video format, so in the future implementation of uncompressed formats may possible as well, so it may support MPEG4 format. Multiple frames embedding are possible. Now we are embedding single frame at a time, but in future multiple frames embedding is also possible.

ACKNOWLEGDEMENT

We would like to express our gratitude towards a number of people whose support and consideration has been an invaluable asset during the course of this work.

REFERENCES

[1] F.A.P. Petitcolas, R.J. Anderson, and M.G. Kuhn, "Information Hiding—A Survey," Proc. IEEE, 1999

[2] Niels Provos and Peter Honeyman, "Hide and Seek: An Introduction to Steganography", University of Michigan, IEEE 2003

[3] Mamta Juneja, Parvinder S. Sandhu, and Ekta Walia,"Application of LSB Based Steganographic Technique for 8-bit Color Images", WASET 2009

[4] Sutaone, M.S.; Khandare, "Image based Steganography using LSB

insertion technique", IET, 2008.

[5] Mazdak Zamani, Azizah A. Manaf, and Shahidan Abdullah, "A Genetic-Algorithm-Based Approach for Audio Steganography" WASET 2009 Digital Information Management, 2006

[7] Kharrazi, M., Sencar, H. T., and Memon, N. (2004). Image steganography: Concepts and practice. In WSPC Lecture Notes Series.

[8] Mobasseri, B.: Direct sequence watermarking of digital video using mframes, Proc. International Conference on Image Processing, Chicago.

[9] R. Chandramouli and N. Memon, "Analysis of lsb based image steganography techniques," Proc. IEEE

International Conf. on Image Processing 3, pp. 1019 –1022, 2001.

[10] F 5 algorithm implementation: 2009, Fridrich, J.R.Du, M. Long: Steganalysis InColor Images, Binghamton, 2007.

IT 014

CLOUD COMPUTING USING VIRTUALIZATION

Name : Priyanka Shivaji Kadam
Sonal B Kutade
Pratik S Kakade
Shailesh V Kamthe

Designation : Student

PDEA's COEManjari (BK), Hadapsar,Pune Maharashtra India

Email ID : priyanka.kadam08@gmail.com Contact no: 9767042128

ABSTRACT

Cloud computing refers a paradigm shift to overall IT solutions while raising the accessibility,Scalability and effectiveness through its enabling technologies. However, migrated cloud platforms and services cost benefits as well as performances are neither clear nor summarized. Globalization and the recessionary economic times have not only raised the bar of a better IT delivery models but also have given access to technology enabled services via internet. Cloud computing has vast potential in terms of lean Retail methodologies that can minimize the operational cost by using the third party based IT capabilities, as a service. It will not only increase the ROI but will also help in lowering the total cost of ownership. In this paper we have tried to compare the cloud computing cost benefits with the actual premise cost which an organization incurs normally. However, in spite of the cost benefits, many IT professional believe that the latest model i.e. “cloud computing” has risks and security concerns. This report demonstrates how to answer the following questions:
(1) Idea behind cloud computing. (2) Monetary cost benefits of using cloud with respect to traditional premise computing. (3) What are the various security issues and how these threats can be mitigated?
We have tried to find out the cost benefit by comparing the Microsoft Azure cloud cost with the prevalent premise cost.

INTRODUCTION

Cloud computing, managed services, software as a service (SaaS), software on demand, Software + Service, platform as a service (PaaS), infrastructure as a service have all been used to describe new ways to build, deliver, and purchase software. Are they just

different names for the same thing, or are they similar names for different ideas? This textbook is written as a primer for anyone trying to make sense of what many believe is the 3rd major wave of computing, after mainframe, and client-server computing. While it contains some aspects of technology the book is first and foremost written from a business perspective. Hence, we open with seven business models for software and use these models in the discussions throughout the remaining chapters. The book is written for any business contemplating developing new application cloud services for internal or external usage, or for those considering moving existing applications to the cloud. In that light we introduce a five-layer cloud services stack to educate you on many of the new cloud services you can buy and not build. Finally, this textbook is not written for the experts in marketing to read the marketing section, or experts in operations to read the operations chapter. Instead, our mission is to deliver a holistic and balanced view useful for the technology student to understand the business is challenges and the business student to understand the technology challenges.

Cloud Computing is a term from the range of the network architecture. The term Cloud Computing describes a concept, which stands in close relationship with Grid Computing technology.Cloud Computing describes a concept, which stands in close relationship with the Grid **Computing technology.** In Simple words Reduce Infrastructure & Improve Application Performance.

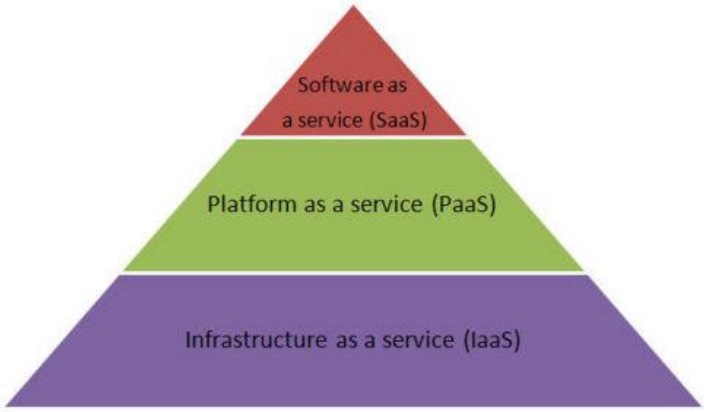
Users of Cloud services operate software applications, platforms and the hardware and/or infrastructure no longer necessary for it, but refer these achievements over Cloud service provider. Applications, platforms, data and infrastructures are not therefore any longer on local systems of the users, but - metaphorically spoken - in a cloud (Cloud) over a number of distant systems distributes. The basis for the access to these distant systems is the Internet, which beside the connection between offers and demanding parties also for connections between instances of the concept of different service tenders (thus different Clouds) makes possible. That access been made by means of defined interfaces as e.g. over a Web browser.

CLOUD COMPUTING FEATURES.

While the lower layers on technological achievements of the Grid Computing construct, Cloud Computing can be understood as similar concept on higher level. Generally three developments are differentiated from Clouds:

* **Software as A service (SaaS)** - providers offer their software in the **Internet as service**. The user has neither knowledge nor control of the infrastructure which is under it.
* **Platform as A service (PaaS)** - providers offer portals or platforms to those the entrance to software services on the one hand to facilitate and on the other hand combinations of services (machine Ups) to make possible are. PaaS is an advancement of the SaaS model. Platforms and portals can affect the paragraph of software services crucially (network effects).
* **Infrastructure as A service (IaaS)** - providers are to be understood as equivalent about SaaS for hardware. IaaS providers offer specific infrastructure services like e.g. Memory or computing services on, which is made available higher levels (SaaS, PaaS). In addition beside SaaS, PaaS and IaaS Cloud markets, on which services are acted, can

be differentiated. This new form of the **IT Oservice-Intermediation** momentarily still is in the development phase.



CLOUD TYPE :

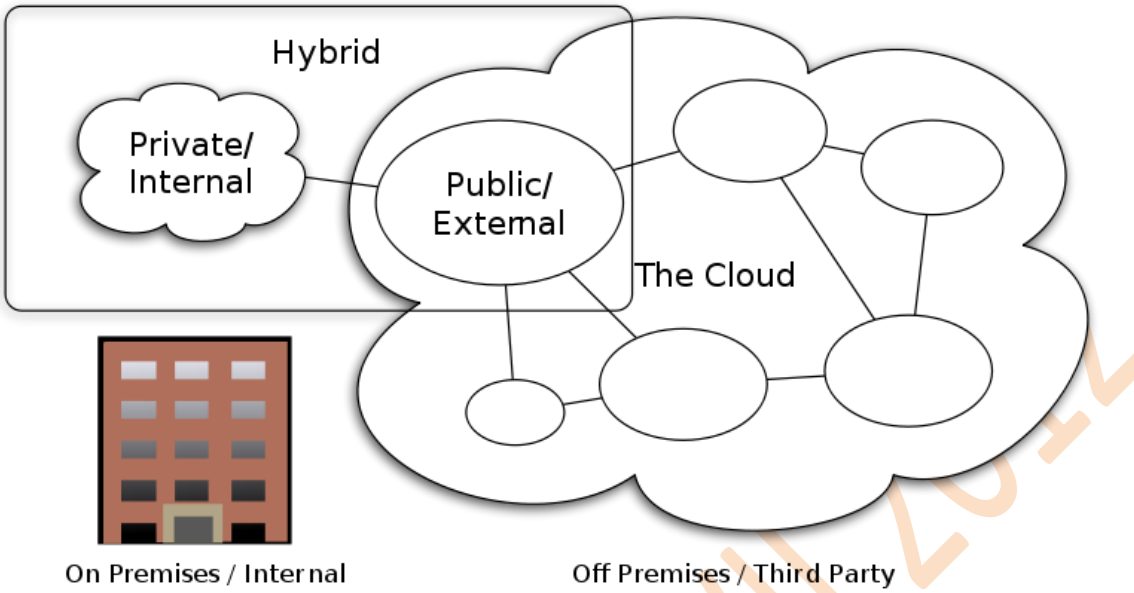
There are three primary deployment models for cloud services:

Private clouds, whether operated and hosted by your enterprise IT department or by an external provider, are for the exclusive use of your organization.

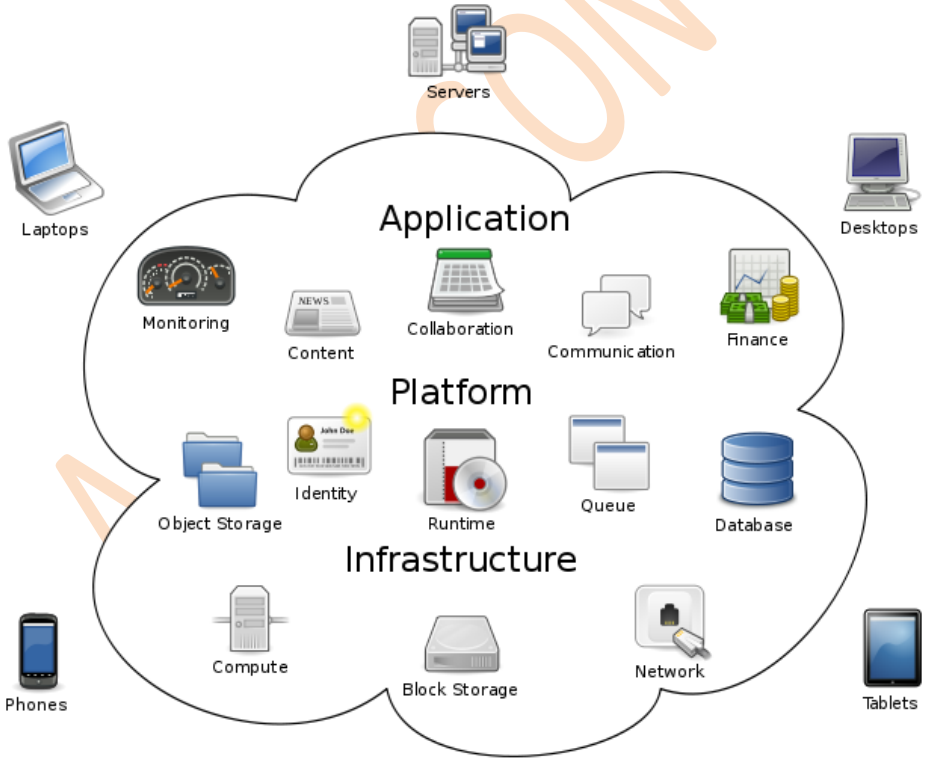
Public clouds are open to any number of organizations and individual users on a shared basis. Using a public cloud minimizes your initial capital investment and combines agility and efficiency with massive scalability.

Hybrid clouds link private and public clouds, providing access to extra resources when the private cloud hits maximum utilization. Or, a hybrid cloud might split computing by tier between private and public clouds. For example, the database may reside in the private cloud while the application server is located in the public cloud.

With any of these structures, cloud computing enables an application to take advantage of idle or excess compute, storage and network capacity that is shared with other applications. The cloud is one of the keys to avoiding overprovisioning and enabling efficient load balancing among your computing resources.



Cloud Computing Types CC-BY-SA 3.0 by Sam Johnston



Cloud Computing

Characteristics

Cloud computing exhibits the following key characteristics:

Empowerment of end-users of computing resources by putting the provisioning of those resources in their own control, as opposed to the control of a centralized IT service (for example)

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

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

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

Device and location independence enable users to access systems using a web browser regardless of their location or what device they are using (e.g., PC, mobile phone). As infrastructure is off-site (typically provided by a third-party) and accessed via the Internet, users can connect from anywhere.

Multi-tenancy enables sharing of resources and costs across a large pool of users thus allowing for: Centralization of infrastructure in locations with lower costs (such as real estate, electricity, etc.)

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

Utilisation and efficiency improvements for systems that are often only 10–20% utilised.

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

Scalability and Elasticity via dynamic ("on-demand") provisioning of resources on a fine-grained, self-service basis near real-time, without users having to engineer for peak loads.

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

Security could improve due to centralization of data, increased security-focused resources, etc., but concerns can persist about loss of control over certain sensitive data, and the lack of security for stored kernels. Security is often as good as or better than other traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford. However, the complexity of security is greatly increased when data is distributed over a wider area or greater number of devices and in multi-tenant systems that are being shared by unrelated users. In addition, user access to security audit logs may be difficult or impossible. Private cloud installations are in part motivated by users' desire to retain control over the infrastructure and avoid losing control of information security.

Maintenance of cloud computing applications is easier, because they do not need to be installed on each user's computer.

Implementing CLOUD

1. Conduct a strategic Diagnostic

The objective of the strategic diagnostic is to identify the major factors influencing the decision to move to the cloud environment and determine the best approach. During the diagnostic step, we will validate the key objectives of moving to the cloud and the "pain points" that the organization wants to address. The drivers during this step include reducing day-to-day risk, reducing the level of involvement in day-to-day IT management, eliminating overhead, achieving better productivity, reducing or eliminating the cost of adding additional users, and protecting the information system from misuse and unauthorized disclosure.

The primary areas to be addressed during the diagnostic step are security and privacy, technical, business and customer impact, economics, and governance and policy. We will evaluate the implications of moving to the cloud environment in each of these categories and document the key issues and considerations revealed during the diagnostic step. The outcome of this diagnostic will be sufficient analysis to support a "go/no go" decision to move to a cloud computing environment and the development of an agreed-on cloud strategy.

2. Define a CLOUD STRATEGY

To define a cloud strategy, the organization should document a complete understanding of each component of its existing architecture. The analysis examines the required user services, processing services, information security, application software standards, and integrated software down to each component. This can be achieved by leveraging existing architectural documents and ensuring the appropriate level of detail is documented, including system-to-system interfaces, data storage, forms processing and reporting, distributed architecture, access control (authentication and authorization), and security and user provisioning.

3. Create an Implementation Plan

The implementation plan identifies the roles and responsibilities, operating model, major milestones, Work Breakdown Structure (WBS), risk plan, dependencies, and quality control mechanisms to implement the cloud strategy successfully.

After completing Phase 1, the organization will have fully analyzed its options, identified all requirements, thoroughly assessed short-term and long-term costs and benefits, gained executive governance approval, and socialized the solution with stakeholders (including oversight entities). This phase ensures that the organization will have a high degree of confidence in successfully moving to the cloud environment, reap the expected benefits, not constrain future functionality, and avoid hidden future costs.

The Cloud Deployment phase (Phase 2) focuses on implementing the strategy developed in the planning phase. Leveraging the various cloud models helps identify the most effective solution(s) based on the existing organization architecture. Some of the criteria used in recommending a vendor are the vendor's primary service model (i.e., infrastructure, platform, or software), business model, how much existing technology can they leverage, end-user experience, and risks involved in porting to the cloud. Deploying to the cloud involves taking the decision analysis from Phase 1 as input and proceeding with the following four step.

Functional Requirement

A functional requirement defines a function of a software system or its component. Functional requirements may be specific functionality that define *what* a system is supposed to accomplish.

HARDWARE AND SOFTWARE REQIREMENTS:

Hardware Used:-

DELL INTEL CORE i7 laptop & LG LAPTOP supporting Intel VT techinology
Operating System :- Windows XP, Windows 2008, UBUNTU Server X64.
4 GB RAM, 500 GB HDD;
Monitor Capable of 800×600 display at 16-bit High color.
24 port network switch.

Software Used:-

VMWARE SERVER
VMWARE ESXi
SQL SERVER 2008
HYPER-V
CITRIX XEN APP

ECONOMIC ASPECTS

In order to allow for economic considerations, cloud systems should help in realising the following aspects:

Cost reduction is one of the first concerns to build up a cloud system that can adapt to changing consumer behaviour and reduce cost for infrastructure maintenance and acquisition. *Scalability* and *Pay per Use* are essential aspects of this issue. Notably, setting up a cloud system typically entails additional costs – be it by adapting the business logic to the cloud host specific interfaces or by enhancing the local infrastructure to be "cloud-ready". See also *return of investment* below.

Pay per use. The capability to build up cost according to the actual consumption of resources is a relevant feature of cloud systems. Pay per use strongly relates to quality of service support, where specific requirements to be met by the system and hence to be paid for can be specified. One of the key economic drivers for the current level of interest in cloud computing is the structural change in this domain. By moving from the usual capital upfront investment model to an operational expense, cloud computing promises to enable especially SME's and entrepreneurs to accelerate the development and adoption of innovative solutions.

Improved time to market is essential in particular for small to medium enterprises that want to sell their services quickly and easily with little delays

caused by acquiring and setting up the infrastructure, in particular in a scope compatible and competitive with larger industries. Larger enterprises need to be able to publish new capabilities with little overhead to remain competitive. Clouds can support this by providing infrastructures, potentially dedicated to specific use cases that take over essential capabilities to support easy provisioning and thus reduce time to market.

Return of investment (ROI) is essential for all investors and cannot always be guaranteed – in fact some cloud systems currently fail this aspect. Employing a cloud system must ensure that the cost and effort vested into it is outweighed by its benefits to be commercially viable – this may entail direct (e.g. more customers) and indirect (e.g. benefits from advertisements) ROI. Outsourcing resources versus increasing the local infrastructure and employing (private) cloud technologies need therefore to be outweighed and critical cut-off points identified.

Turning CAPEX into OPEX is an implicit, and much argued characteristic of cloud systems, as the actual cost benefit (cf. ROI) is not always clear (see e.g.[9]). Capital expenditure (CAPEX) is required to build up a local infrastructure, but with outsourcing computational resources to cloud systems on demand and scalable, a company will actually spend operational expenditure (OPEX) for provisioning of its capabilities, as it will acquire and use the resources according to operational need.

"Going Green" is relevant not only to reduce additional costs of energy consumption, but also to reduce the carbon footprint. Whilst carbon emission by individual machines can be quite well estimated, this information is actually taken little into consideration when scaling systems up. Clouds principally allow reducing the consumption of unused resources (down-scaling). In addition, up-scaling should be carefully balanced not only with cost, but also carbon emission issues. Note that beyond software stack aspects, plenty of Green IT issues are subject to development on the hardware level.

TECHNOLOGICAL ASPECTS

The main technological challenges that can be identified and that are commonly associated with cloud systems are:

Virtualization is an essential technological characteristic of clouds which hides the technological complexity from the user and enables enhanced flexibility (through aggregation, routing and translation).

More concretely, virtualization supports the following features: **Ease of use***:* through hiding the complexity of the infrastructure (including management, configuration etc.) virtualization can make it easier for the user to develop new applications, as well as reduces the overhead for controlling the system.

Infrastructure independency*:* in principle, virtualization allows for higher interoperability by making the Code platform independent.

Flexibility and Adaptability*:* by exposing a virtual execution environment, the underlying infrastructure can change more flexible according to different conditions and requirements (assigning more resources, etc.).

Location independence*:* services can be accessed independent of the physical location of the user and the resource.

Multi-tenancy is a highly essential issue in cloud systems, where the location of code and / or data is principally unknown and the same resource may be assigned to multiple users (potentially at the same time). This affects infrastructure resources as well as data / applications / services that are hosted on shared resources but need to be made available in multiple isolated instances. Classically, all information is maintained in separate databases or tables, yet in more complicated cases information may be concurrently altered, even though maintained for isolated tenants. Multitenancy implies a lot of potential issues, ranging from data protection to legislator issues.

S

ecurity, Privacy and Compliance is obviously essential in all systems dealing with potentially sensitive data and code.

Data Management is an essential aspect in particular for storage clouds, where data is flexibly distributed across multiple resources. Implicitly, data consistency needs to be maintained over a wide distribution of *replicated* data sources. At the same time, the system always needs to be aware of the data location (when replicating across data centres) taking latencies and particularly workload into consideration. As size of data may change at any time, data management addresses both horizontal and vertical aspects of scalability. Another crucial aspect of data management is the provided consistency guarantees (eventual vs. strong consistency, transactional isolation vs. no isolation, atomic operations over individual data items vs. multiple data times etc.).

APIs and / or Programming Enhancements are essential to exploit the cloud features: common Programming models require that the developer takes care of the scalability and autonomic capabilities him / herself, whilst a cloud environment provides the features in a fashion that allows user to leave such management to the system.

Metering of any kind of resource and service consumption is essential in order to offer elastic pricing, charging and billing. It is therefore a pre-condition for the elasticity of clouds.

Tools are generally necessary to support development, adaptation and usage of cloud services.

REFERENCE:

1 "Gartner Says Cloud Computing Will Be As Influential As E-business". Gartner.com. Retrieved 2010-08-22.
2 Gruman, Galen (2008-04-07). "What cloud computing really means". *InfoWorld*. Retrieved 2009-06-02.
3 "Cloud Computing: Clash of the clouds". The Economist. 2009-10-15. Retrieved 2009-11-03.
4 Cloud Computing Defined 17 July 2010. Retrieved 26 July 2010.
5 "Kerravala, Zeus, Yankee Group, "Migrating to the cloud is dependent on a converged infrastructure," Tech Target". Convergedinfrastructure.com. Retrieved 2011-12-02.
6 "Baburajan, Rajani, "The Rising Cloud Storage Market Opportunity Strengthens Vendors," infoTECH, August 24, 2011". It.tmcnet.com. 2011-08-24. Retrieved 2011-12-02.
7 "Oestreich, Ken, "Converged Infrastructure," CTO Forum, November 15, 2010". Thectoforum.com. 2010-11-15. Retrieved 2011-12-02.
8 Buyya, Rajkumar; Chee Shin Yeo, Srikumar Venugopal (PDF). *Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities*. Department of Computer Science and Software Engineering, University of Melbourne, Australia. p. 9. Retrieved 2008-07-31.
9 Lillington, Karlin. "Getting clear about cloud computing". *The Irish Times*.
10 Thomas J. Kwasniewski, EJ Puig, "Cloud Computing in the Government", *Data & Analysis Centre for Software*, July 2011
11 "What's In A Name? Utility vs. Cloud vs Grid". Datacenterknowledge.com. Retrieved 2010-08-22.
12 "Distributed Application Architecture". Sun Microsystem. Retrieved 2009-06-16.
13 "Sun CTO: Cloud computing is like the mainframe". Itknowledgeexchange.techtarget.com. 2009-03-11. Retrieved 2010-08-22.
14 "It's probable that you've misunderstood 'Cloud Computing' until now". TechPluto. Retrieved 2010-09-14.
15 "Recession Is Good For Cloud Computing – Microsoft Agrees". CloudAve.
16 "Defining "Cloud Services" and "Cloud Computing"". IDC. 2008-09-23. Retrieved 2010-08-22.

IT 015

IMAGE IDENTIFICATION USING CBIR

#Suvarna Vitthal Khandagale,Sweety Kamthe,Rajashree Salunkhe
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id-suvarnakhandagale30@gmail.com

Contact no.9730462558

*Annu Shabbir Tamboli
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id-tamboliannu@yahoo.in

Contact no.9096233439

Abstract

There is growing interest in CBIR because of the limitations inherent in metadata-based systems, as well as the large range of possible uses for efficient image retrieval. Textual information about images can be easily searched using existing technology, but requires humans to personally describe every image in the database. This is impractical for very large databases, or for images that are generated automatically, e.g. from surveillance cameras. It is also possible to miss images that use different synonyms in their descriptions. Systems based on categorizing images in semantic classes like "cat" as a subclass of "animal" avoid this problem but still face the same scaling issues. Our aim is to to build an application depending on Content-based image retrieval (CBIR). Our main aim is to filter the images and to retrieve the images that contains the data as per the query provided to the application. Secondary aim is to have this application utilized in law enforcement regarding access to the images. For examples we could make use of this application to keep a check on which images should be accessible to small children.

Keywords-CBIR, Color Histogram, Image retrieval, NNS, SIFT.

INTRODUCTION

The purpose of the project is to develop a CBIR based applications that could help in the filtering the images so as to get almost 100% accurate search results. Also using this application we could have law enforced for having access to the images.

Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases.

"Content-based" means that the search will analyze the actual contents of the image. The term 'content' in this context might refer colors, shapes, textures, or any other information that can be derived from the image itself. Without the ability to examine image content, searches must rely on metadata such as captions or keywords, which may be laborious or expensive to produce.

Innovativeness: Till now we have search engines or application that perform search operation based on the query we send to it as input and match it with the names of the entity or the text. So we are able to extract only those entities which have proper naming i.e. name as per data or entity the file contains. With this application we are able to extract entities or files based on the data it actually contains. So, we do not have to depend on the naming system to extract exact data.This would more effective in case of searching images as they are the ones which difficult to extract if not properly named.

TECHNIQUES USED IN CBIR

CBIR operates on a totally different principle, retrieving/searching stored images from a collection by comparing features automatically extracted from the images themselves. The commonest features used are mathematical measures of color, texture or shape (basic). A system (CBIR) allows users to formulate queries by submitting an example of the type of image being sought (input), though some offer alternatives such as selection from a palette or sketch input we can also select color textures or any other visual information. The system then identifies those stored images whose feature values match those of the query most closely (right side), and displays thumbnails of these images on the screen.

A. Colour

One of the most important features that make possible the recognition of images by humans is colour. Colour is a property that depends on the reflection of light

to the eye and the processing of that information in the brain. We use colour everyday to tell the difference between objects, places, and the time of day [7]. Usually colours are defined in three dimensional colour spaces. These could either be ***RGB*** (Red, Green, and Blue), ***HSV*** (Hue, Saturation, and Value) or ***HSB*** (Hue, Saturation, and Brightness). The last two are dependent on the human perception of hue, saturation, and brightness.

Most image formats such as ***JPEG***, ***BMP***, ***GIF***, use the RGB colour space to store information [7]. The RGB colour space is defined as a unit cube with red, green, and blue axes. Thus, a vector with three co-ordinates represents the colour in this space. When all three coordinates are set to zero the colour perceived is black. When all three coordinates are set to 1 the colour perceived is white [7]. The other colour spaces operate in a similar fashion but with a different perception.

Methods of Representation

The main method of representing colour information of images in CBIR systems is through colour histograms. A colour histogram is a type of bar graph, where each bar represents a particular colour of the colour space being used. In MatLab for example you can get a colour histogram of an image in the RGB or HSV colour space. The bars in a colour histogram are referred to as bins and they represent the x-axis. The number of bins depends on the number of colours there are in an image. The y-axis denotes the number of pixels there are in each bin. In other words how many pixels in an image are of a particular colour.

An example of a color histogram in the HSV color space can be seen with the following image:

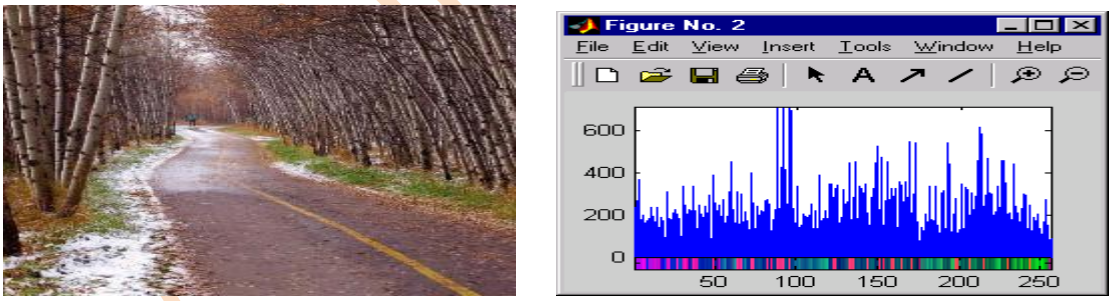


Fig. 1 Sample Image and its Corresponding Histogram

To view a histogram numerically one has to look at the color map or the numeric representation of each bin.

As one can see from the color map each row represents the color of a bin. The row is composed of the three coordinates of the color space. The first coordinate represents hue, the second saturation, and the third, value, thereby giving HSV. The percentages of each of these coordinates are what make up the color of a

bin. Also one can see the corresponding pixel numbers for each bin, which are denoted by the blue lines in the histogram.

Quantization in terms of color histograms refers to the process of reducing the number of bins by taking colors that are very similar to each other and putting them in the same bin. By default the maximum number of bins one can obtain using the histogram function in MatLab is 256. For the purpose of saving time when trying to compare color histograms, one can quantize the number of bins. Obviously quantization reduces the information regarding the content of images but as was mentioned this is the tradeoff when one wants to reduce processing time.

Color Map (x-axis)			***Number of Pixels per Bin (y-axis)***
H	***S***	***V***	
0.9922	0.9882	0.9961	106
0.9569	0.9569	0.9882	242
0.9725	0.9647	0.9765	273
0.9176	0.9137	0.9569	372
0.9098	0.8980	0.9176	185
0.9569	0.9255	0.9412	204
0.9020	0.8627	0.8980	135
0.9020	0.8431	0.8510	166
0.9098	0.8196	0.8078	179
0.854	0.85	0.894	188

9	10	1	
0.8235	0.8235	0.8941	241
0.8471	0.8353	0.8549	104
0.8353	0.7961	0.8392	198
.	.	.	.
.	.	.	.
.	.	.	.

TABLE I COLOR MAP AND NUMBER OF PIXELS FOR THE PREVIOUS IMAGE

There are two types of colour histograms, Global colour histograms (***GCH***s) and Local colour histograms (***LCH***s). A GCH represents one whole image with a single colour histogram. An LCH divides an image into fixed blocks and takes the colour histogram of each of those blocks [7]. LCHs contain more information about an image but are computationally expensive when comparing images. "The GCH is the traditional method for colour based image retrieval. However, it does not include information concerning the colour distribution of the regions [7]" of an image. Thus when comparing GCHs one might not always get a proper result in terms of similarity of images.

B. Texture

Texture is that innate property of all surfaces that describes visual patterns, each having properties of homogeneity. It contains important information about the structural arrangement of the surface, such as; clouds, leaves, bricks, fabric, etc. It also describes the relationship of the surface to the surrounding environment [2]. In short, it is a feature that describes the distinctive physical composition of a surface.

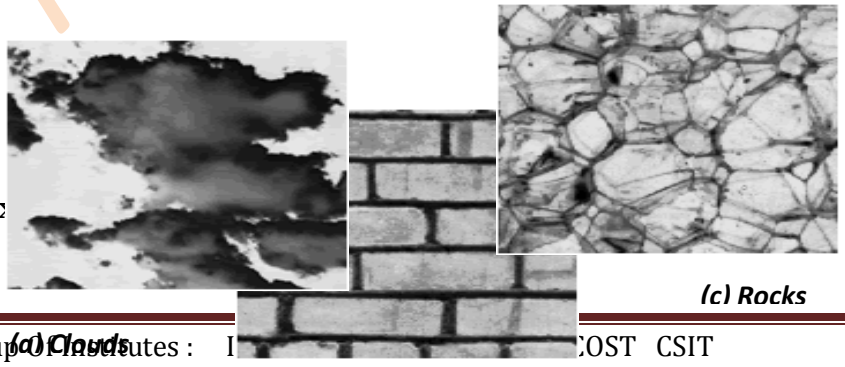


Fig. 2 Ex

(a) Clouds *(b) Bricks* *(c) Rocks*

Texture properties include:

- Coarseness
- Contrast
- Directionality
- Line-likeness
- Regularity
- Roughness

Texture is one of the most important defining features of an image. It is characterized by the spatial distribution of gray levels in a neighborhood [8]. In order to capture the spatial dependence of gray-level values, which contribute to the perception of texture, a two-dimensional dependence texture analysis matrix is taken into consideration. This two-dimensional matrix is obtained by decoding the image file; jpeg, bmp, etc.

Methods of Representation

There are three principal approaches used to describe texture; statistical, structural and spectral.

- Statistical techniques characterize textures using the statistical properties of the grey levels of the points/pixels comprising a surface image. Typically, these properties are computed using: the grey level co-occurrence matrix of the surface, or the wavelet transformation of the surface.
- Structural techniques characterize textures as being composed of simple primitive structures called “texels” (or texture elements). These are arranged regularly on a surface according to some surface arrangement rules.
- Spectral techniques are based on properties of the Fourier spectrum and describe global periodicity of the grey levels of a surface by identifying high-energy peaks in the Fourier spectrum[9] .

For optimum classification purposes, what concern us are the statistical techniques of characterization… This is because it is these techniques that result in computing texture properties… The most popular statistical representations of texture are:

- Co-occurrence Matrix
- Tamura Texture
- Wavelet Transform

Co-occurrence Matrix

Originally proposed by R.M. Haralick, the co-occurrence matrix representation of texture features explores the grey level spatial dependence of texture [2]. A mathematical definition of the co-occurrence matrix is as follows [4]:

-Given a position operator ***P(i,j)***,
-let ***A*** be an n x n matrix
-Whose element ***A[i][j]*** is the number of times that points with grey level (intensity) ***g[i]*** occur, in the position specified by ***P***, relative to points with grey level ***g[j]***.
-Let ***C*** be the ***n x n*** matrix that is produced by dividing ***A*** with the total number of point pairs that satisfy ***P***. ***C[i][j]*** is a measure of the joint probability that a pair of points satisfying ***P*** will have values ***g[i]***, ***g[j]***.
-***C*** is called a co-occurrence matrix defined by ***P***.
Examples for the operator ***P*** are: "***i*** above ***j***", or "***i*** one position to the right and two below ***j***", etc.

This can also be illustrated as follows… Let *t* be a translation, then a co-occurrence matrix C_t of a region is defined for every grey-level *(a, b)* by :

$$C_t(a,b) = card\{(s, s+t) \in R^2 \mid A[s] = a, A[s+t] = b\}$$

Here, $C_t(a, b)$ is the number of site-couples, denoted by *(s, s + t)* that are separated by a translation vector *t*, with *a* being the grey-level of *s*, and *b* being the grey-level of *s + t*.

For example; with an 8 grey-level image representation and a vector *t* that considers only one neighbor, we would find:

1	2	1	3	4
2	3	1	2	4
3	3	2	1	1

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	0	0	0	0	0
2	0	1	0	2	0	0	0	0
3	0	0	1	1	0	0	0	0
4	0	1	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Fig. 3 Classical Co-occurrence matrix

At first the co-occurrence matrix is constructed, based on the orientation and distance between image pixels[2]. Then meaningful statistics are extracted from the matrix as the texture representation. Haralick proposed the different texture features[10].For each Haralick texture feature, we obtain a co-occurrence matrix. These co-occurrence matrices represent the spatial distribution and the dependence of the grey levels within a local area. Each ***(i,j)*** th entry in the matrices, represents the probability of going from one pixel with a grey level of '***i***' to another with a grey level of '***j***' under a predefined distance and angle. From these matrices, sets of statistical measures are computed, called feature vectors[11] .

Tamura Texture

By observing psychological studies in the human visual perception, Tamura explored the texture representation using computational approximations to the three main texture features of: coarseness, contrast, and directionality[2,12]. Each of these texture features are approximately computed using algorithms…

- *Coarseness* is the measure of granularity of an image[12] , or average size of regions that have the same intensity [13].
- *Contrast* is the measure of vividness of the texture pattern. Therefore, the bigger the blocks that makes up the image, the higher the contrast. It is affected by the use of varying black and white intensities[12].
- *Directionality* is the measure of directions of the grey values within the image[12] .

Wavelet Transform

Textures can be modeled as quasi-periodic patterns with spatial/frequency representation. The wavelet transform transforms the image into a multi-scale representation with both spatial and frequency characteristics. This allows for effective multi-scale image analysis with lower computational cost [2]. According to this transformation, a function, which can represent an image, a curve, signal etc., can be described in terms of a coarse level description in addition to others with details that range from broad to narrow scales [11].

Unlike the usage of sine functions to represent signals in Fourier transforms, in wavelet transform, we use functions known as wavelets. Wavelets are finite in time, yet the average value of a wavelet is zero [2]. In a sense, a wavelet is a waveform that is bounded in both frequency and duration. While the Fourier

transform converts a signal into a continuous series of sine waves, each of which is of constant frequency and amplitude and of infinite duration, most real-world signals (such as music or images) have a finite duration and abrupt changes in frequency. This accounts for the efficiency of wavelet transforms. This is because wavelet transforms convert a signal into a series of wavelets, which can be stored more efficiently due to finite time, and can be constructed with rough edges, thereby better approximating real-world signals [14].

Examples of wavelets are Coiflet, Morlet, Mexican Hat, Haar and Daubechies. Of these, Haar is the simplest and most widely used, while Daubechies have fractal structures and are vital for current wavelet applications [2]. These two are outlined below:

Haar Wavelet

The Haar wavelet family is defined as [2]:

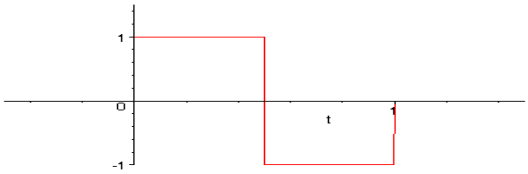


Fig. 4 Haar Wavelet Example

Daubechies Wavelet

The Daubechies wavelet family is defined as[2] :

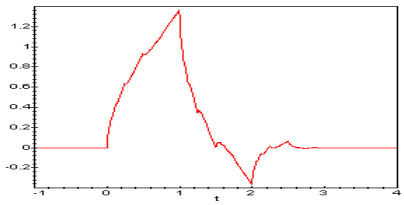


Fig. 5 Daubechies Wavelet Example

C. Shape[14]

Shape may be defined as the characteristic surface configuration of an object; an outline or contour. It permits an object to be distinguished from its surroundings by its outline[15] . Shape representations can be generally divided into two categories[2] :

- Boundary-based, and
- Region-based.

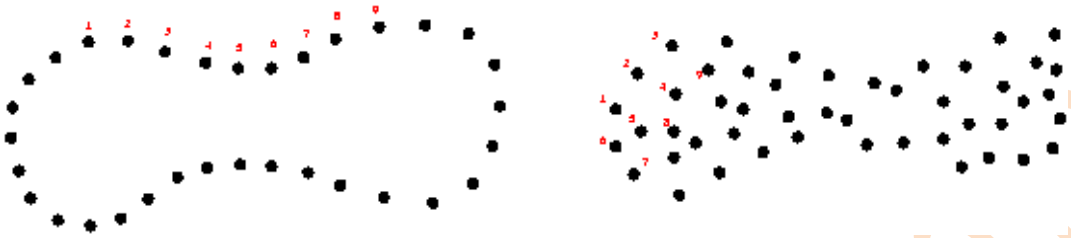


Fig. 6 Boundary-based & Region-based

Boundary-based shape representation only uses the outer boundary of the shape. This is done by describing the considered region using its external characteristics; i.e., the pixels along the object boundary. Region-based shape representation uses the entire shape region by describing the considered region using its internal characteristics; i.e., the pixels contained in that region [17].

Methods of Representation

For representing shape features mathematically, we have[16] :

Boundary-based:

- Polygonal Models, boundary partitioning
- Fourier Descriptors
- Splines, higher order constructs
- Curvature Models

Region-based:

- Super quadrics
- Fourier Descriptors
- Implicit Polynomials
- Blum's skeletons

The most successful representations for shape categories are Fourier Descriptor and Moment Invariants[2]:

- The main idea of Fourier Descriptor is to use the Fourier transformed boundary as the shape feature.
- The main idea of Moment invariants is to use region-based moments, which are invariant to transformations as the shape feature.

ALGORITHMS

A. The SIFT (Scale-Invariant Feature Transform) algorithm[15]

The SIFT algorithm identifies features of an image that are distinct, and these features can in turn be used to identify similar or identical objects in other images. We will here give an introduction to the SIFT algorithm.

SIFT has four computational phases. The reason for this being that some computations performed by SIFT are very expensive. The cost of extracting the keypoints is minimized by the cascading approach of SIFT. The more expensive operations are only applied on locations that pass an initial, cheaper test. The output of the SIFT algorithm is a set of keypoint descriptors2. Once such descriptors have been generated for more than one image, one can begin image matching.

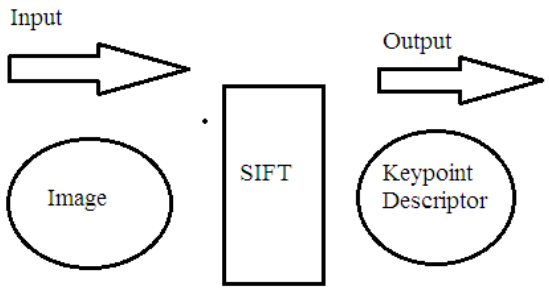


Fig. 7 Scale-Invariant Feature Transform

SIFT takes as input an image, and generates a set of keypoints descriptors. The keypoints descriptors may then be stored in a separate file.

The image matching, or object matching, is not part of the SIFT algorithm. For matching we use a nearest neighbor search (NNS), an algorithm that is able to detect similarities between keypoints. Thus, SIFT only makes matching possible by generating the keypoints descriptors.

B. Nearest neighbor search[15]

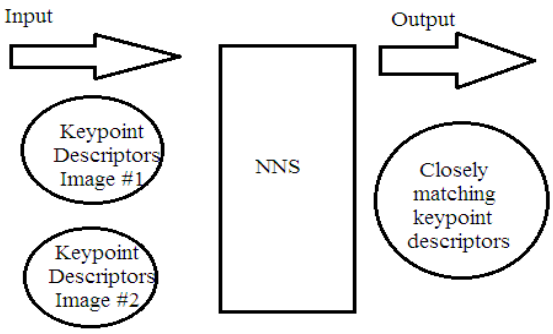


Fig. 8 Nearest neighbor search

When we check for an image match, the two sets of keypoints descriptors are given as input to a nearest neighbor search algorithm. The output of the algorithm is a set of keypoints descriptors found to be very similar.

Keypoints matching

As mentioned earlier, this is accomplished by a NNS in a kd-Tree. However, though the matching utilizes a NNS, the criterion for a match is not to be a nearest neighbor; then all nodes would have a match. Note, as stated before, the comparison is done from the source image, to the compared image. Thus, in the following outline, to "select a node" means to select a node from

the keypoints of the source image. The procedure is as follows:

1. Select a node from the set of all nodes not yet selected.

2. Mark the node as selected.

3. Locate the two nearest neighbors of the selected node.

4. If the distance between the two neighbors are less than or equal to a

given distance, we have a match. Mark the keypoints as match.

5. Perform step 1-4 for all nodes in the source image.

The key step of the matching algorithm is step 4. It is here it is decided whether the source node (or keypoints) has a match in the compared set of keypoints, or not.

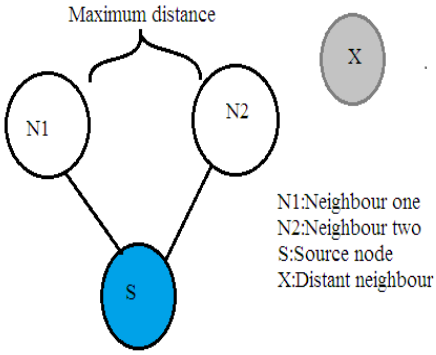


Fig. 9 Step of matching procedure

The image matching algorithm verifies the distance between the two closest neighbors of a source node. If the distance is within a given boundary, the source node is marked as a match.

Quality of Match formula[15]

Let *Ks* be the number of keypoints in the source image, *Kc* be the number of keypoints in the compared image, and *Km* be the number of matching keypoints.

We can rewrite the QoM formula to-

$$\mathbf{QoM}= \frac{Km*Kc}{Ks^2} * 100$$

Ks » Kc = very unreliable matching

Ks « Kc = very reliable matching

SYSTEM ARCHITECTURE

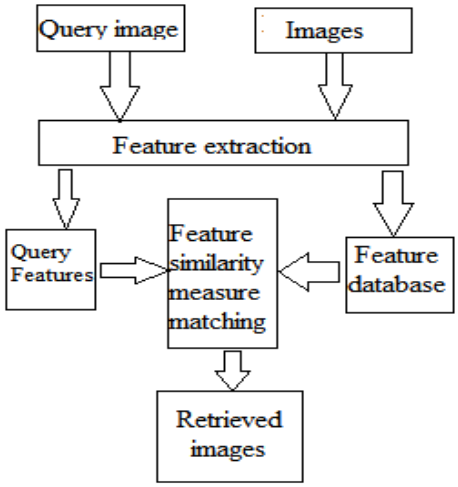


Fig. 10 Architecture of a typical CBIR system

For each image in the image database, its features are extracted and the obtained feature space (or vector) is stored in the feature database. When a query image comes in, its feature space will be compared with those in the feature database one by one and the similar images with the smallest feature distance will be retrieved.

CONCLUSION

The application performs a simple colour-based search in an image database for an input query image, using colour histograms. It then compares the colour histograms of different images using the *Quadratic Distance Equation*. Further enhancing the search, the application performs a texture-based search in the colour results, using wavelet decomposition and energy level calculation. It then compares the texture features obtained using the *Euclidean Distance Equation*. A more detailed step would further enhance these texture results, using a shape-based search.

CBIR is still a developing science. As image compression, digital image processing, and image feature extraction techniques become more developed, CBIR maintains a steady pace of development in the research field. Furthermore, the development of powerful processing power, and faster and cheaper memories contribute heavily to CBIR development. This development promises an immense range of future applications using CBIR.

REFERENCES

Barbeau Jerome, Vignes-Lebbe Regine, and Stamon Georges, "A Signature based on Delaunay Graph and Co-occurrence Matrix," Laboratoire Informatique et Systematique, Universiyt of Paris, Paris, France, July 2002, Found at: http://www.math-info.univ-paris5.fr/sip-lab/barbeau/barbeau.pdf

Sharmin Siddique, "A Wavelet Based Technique for Analysis and Classification of Texture Images," Carleton University, Ottawa, Canada, Proj. Rep. 70.593, April 2002.

Thomas Seidl and Hans-Peter Kriegel, "Efficient User-Adaptable Similarity Search in Large Multimedia Databases," in Proceedings of the 23rd International Conference on Very Large Data Bases VLDB'97, Athens, Greece, August 1997, Found at:
http://www.vldb.org/conf/1997/P506.PDF

FOLDOC, *Free On-Line Dictionary Of Computing*, "cooccurrence matrix," May 1995, [Online Document], Available at:
http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?cooccurrence+matrix

Colin C. Venteres and Dr. Matthew Cooper, "A Review of Content-Based Image Retrieval Systems", [Online Document], Available at:
http://www.jtap.ac.uk/reports/htm/jtap-054.html

Linda G. Shapiro, and George C. Stockman, *Computer Vision*, Prentice Hall, 2001.
Shengjiu Wang, "A Robust CBIR Approach Using Local Color Histograms," Department of Computer Science, University of Alberta, Edmonton, Alberta, Canada, Tech. Rep. TR 01-13, October 2001, Found at:
http://citeseer.nj.nec.com/wang01robust.html

R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision*, McGraw Hill International Editions, 1995.
FOLDOC, *Free On-Line Dictionary Of Computing*, "texture," May 1995, [Online Document], Available at:
http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=texture

"Texture," class notes for *Computerized Image Analysis MN2*, Centre for Image Analysis, Uppsala, Sweden, Winter 2002, Found at:
http://www.cb.uu.se/~ingela/Teaching/ImageAnalysis/Texture2002.pdf

G. D. Magoulas, S. A. Karkanis, D. A. Karras and M. N. Vrahatis, "Comparison Study of Textural Descriptors for Training Neural Network Classifiers", in Proceedings of the 3rd IEEE-IMACS World Multi-conference on Circuits, Systems, Communications and Computers, vol. 1, 6221-6226, Athens, Greece, July 1999, Found at:

http://www.brunel.ac.uk/~csstgdm/622.pdf
Pravi Techasith, "Image Search Engine," Imperial College, London, UK, Proj. Rep., July 2002, Found at:
http://km.doc.ic.ac.uk/pr-p.techasith-2002/Docs/OSE.doc

Bjorn Johansson, "QBIC (Query By Image Content)", November 2002, [Online Document], Available at:
http://www.isy.liu.se/cvl/Projects/VISIT
0bjojo/survey/surveyonCBIR/node26.html

Content Based Image Retrieval By Rami Al-Tayeche (237262) & Ahmed Khalil (296918)

An evaluation of the SIFT algorithm for CBIR by Thomas Bakken

IT 016

FUTURE OF HAP IN POST 3G SYSTEMS

Mr. Nitish Dixit

Student,

Fourth year(Department of Computer Engineering),

Bharati Vidyapeeth Deemed University,

College of Engineering,

Pune-411043. (Maharashtra) INDIA

dixit.nitish@gmail.com

nitish_dixit@rocketmail.com

+91 9765896164

Mr. P.D. Joshi

Asst. Professor,

Department of Computer Engineering,

Bharati Vidyapeeth Deemed University,

College of Engineering,

Pune-411043. (Maharashtra) INDIA

pdprashant@gmail.com

pdjoshi@bvucoep.edu.in

+91 9657702589

Abstract— *5G Technology stands for 5th Generation Mobile technology. The 5G technology network is aimed at offering enhanced and available connectivity around the world. This requires a strong and efficient communication medium which encapsulates very high data transfer rate, easy management and cost efficient hardware .The existing satellite communication medium is not efficient to completely accomplish the above stated goals. In this paper, we focus on a new technology named H.A.P. (High altitude platform) station which is a candidate of post-3G (5G) wireless communication systems, aiming at creating a new broadcast and communications integrated infrastructure to support future generation technologies.*

Keywords: ATM(Asynchronous transfer mode), HAP(High altitude platform), HDTV(High definition telivision), LEO(Low elliptical orbit), UAV(Unmanned aerial vehicle), QOS(quality of service).

1.0 INTRODUCTION:

5G technology is going to be a new mobile revolution in mobile market. 5G technology has extraordinary data capabilities and has ability to tie together unrestricted call volumes and infinite data broadcast within latest mobile operating system. 4G is being developed to accommodate the QOS(Quality of service) and rate requirements set by forthcoming applications like wireless broadband access, Multimedia Messaging Service (MMS), video chat, mobile TV, HDTV content, Digital Video Broadcasting (DVB), minimal services like voice and data, and other services that utilize bandwidth.[3]

The 5th wireless mobile multimedia internet networks are real wireless world, which are completed wireless communication without limitation.

Key concepts suggested in development of 5G and beyond 4G wireless communications are:

Real wireless world with no more limitation with access and zone issues. Wearable devices with AI capabilities.

Internet protocol version 6 (IPv6), where a visiting care-of mobile IP address is assigned according to location and connected network. One unified global standard.

The user can simultaneously be connected to several wireless access technologies and seamlessly move between them. These access technologies can be a 2.5G, 3G, 4G, or 5G mobile networks, Wi-Fi, WPAN, or any other future access technology.

Smart-radio technology: allowing different radio technologies to share the same spectrum efficiently by adaptively finding unused spectrum and adapting the transmission scheme to the requirements of the technologies currently sharing the spectrum.[7] High altitude stratospheric platform station (HAPS) systems.

2.0 HAPS:

High-Altitude Platform Station (HAPS) is one of the promising wireless communication infrastructures in the future. Non-geostationary mobile satellite systems integrated with High Altitude Platforms (HAPs) may have great potential in the next generation telecommunication services.

HAPS is a promising infrastructure for providing various wireless services such as mobile and fixed services in the global area with some advantages of both terrestrial and satellite networks.

HAPs are based on airships or balloons placed at about 20 km height and they combine both the advantages of terrestrial networks (the highest mast in the town) and of satellites (a LEO in very low orbit). As earth observation (EO) satellites need the download of huge amount of data, HAPs could establish an optical link with a LEO satellite in order to exploit the high data rate.

The integration of a satellite system with a HAPS segment appears very suitable to provide communication services, including Internet access, for a large set of applications. In fact, the satellite capability to provide wide coverage and broadband access can be enhanced by the use of cost-effective, mobile/portable and low-power consuming user terminals, when HAPS acts as an intermediate repeater.[4]

Thus the goal of HAP based network is to cover as wide area as possible with deployment of multiple UAVs(unmanned aerial vehicles) , i.e. ultimate goal of HAP network is to deploy as many MBSs to cover dedicated area in order to construct a network structure. In this configuration, UAVs can act as mobile base stations for the network. [12]

An information system formed by HAP (High Altitude Platform) will be a new generation-system for the wireless communications and HAPS (HAP Station) communication system combines the advantages of both terrestrial and satellite communication systems and avoids, to different extents, their disadvantages. It has been identified as a promising communication style in the field of telecommunication and an effective complement to the existing communication system.

Let's take a look at HDTV, which is now a hot topic in the consumer electronic space, and while content can be readily delivered from the studio and other specific locations, it is still quite difficult to deliver live content at short notice

from outside broadcast locations. An uncompressed HDTV signal is preferred by broadcasters for pre broadcast content since compression introduces excessive delay, and if the compression is incurring loss, it is liable to introduce progressive degradation of the signal. The data rate required to deliver an uncompressed video stream is up to 3.0Gbps [1], which is substantially higher than the speed of 10–30 Mbps to transmit 1080i

compressed signal [2]. Currently, it is quite difficult to deliver HDTV pre broadcast content due to the high data rates involved. Using gigabit links from a high-altitude platform (HAP) will provide one possible solution to overcome these delivery pre broadcast material problems, at least for the lower resolution formats of HDTV. [9]

HAP can also offer significant benefits, is to moving objects such as trains. They would use a sophisticated electronically steer able aerial to track the HAPs and would allow continuous reception of signals – even between HAPs.

2.1 TECHNICAL ASPECTS:

To provide higher data rate and QOS guarantee to mobile multimedia users, we propose an ATM (Asynchronous transfer mode) - based broadband wireless communication system. In order to show the feasibility and investigate the key issues of such an ATM-based HAPS[6], there is also a developed a demonstration system based on a 15m X 4.3m remote-controlled airship. By installing a remote-controlled video camera, a wireless HUB, and a transmitter of digital TV in this airship, we have a broadband integrated information system which includes

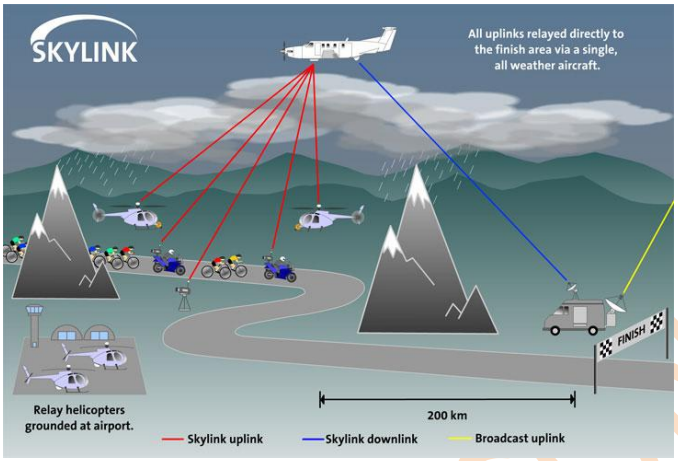
(1) remote sensing of the ground.

(2) broadband wireless Internet access, and

(3) mobile multimedia communication with QOS control.

The link between HAP and ground can be exploited by the use of high data rate RF links in Ka-band or X-band (mainly used in EO satellites)[8], V-band (allocated for HAPs) or W-Band (the new frontier).[5]

Fig 1. Shows a general working and structure of a HAP with the help of skylink.



The above fig.1 shows that for geographically dispersed events such as cycling and marathons, video links are traditionally relayed via multiple helicopters. Skylink replaces all of these low altitude relay platforms with a single high altitude platform; a Pilatus PC-12 cruising at up to 30,000ft. All of the uplinks are relayed via just one aircraft.[2]

2.2 ALTITUDE SELECTION FOR HAP:

The altitude is selected by keeping the following factors in consideration:

Wind Factor which varies with height, weather and location.

Air-traffic due to commercial flights which have to maintain a safe distance from every station installed. Various permissions from different authorities also have to be arranged beforehand.

3.0 COMPARISON WITH SATELLITES AND TERRESTRIAL COMMUNICATION:

HAPS have the following advantages in comparison to terrestrial and satellite mode of communication:

Table 1: System characteristics of HAP, Terrestrial and Satellite systems.[10]

Subject	*HAPs*	*Terrestrial*	*Satellite*
Cell Radius	*3~7 km*	*0.1~2 km*	*50 km for LEO*
BS Coverage area radius	*Typical 30 km. ITU has suggested 150 km.*	*5 km.*	*A few hundred km for LEO*
Elevation angles	*High*	*Low*	*High*
Propagation delay	*Low*	*Low*	*Noticeable*
Propagation Characteristics	*Nearly free space path loss (FSPL)*	*Well established typically Non FPSL*	*FPSL with rain*
BS power supply	*Fuel(ideally solar)*	*Electricity*	*Solar*
BS maintenance	*Less complexity in terms of coverage area*	*Complex if multiple BSs needed to update*	*Impossible*
BS cost	*No specific number but supposed to be economical in terms of*	*Well established market, cost dependent on companies*	*5 billion for Iridium, very expensive*

	coverage area		
Operational cost	*Medium(mainly airship maintenance)*	*Medium ~ High in terms of the number of BSs*	*High*
Deployment complexity	*Low(especially in remote and high density population areas)*	*Medium(more complex to deploy in the city area)*	*High*

4.0 ADVANTAGES OF HAP IN 4G AND POST 4G SYSTEMS

4G/5G services delivered via HAPS are expected

to have the following advantages:

Can be deployed to serve as the macro cell component of the tower-based cells, thus offering a cost effective solution for provision of pico /micro/macro cellular architecture based on a single air interface standard.

Ease the restrictions currently imposed on site availability compared to terrestrial.

More environment-friendly than currently used terrestrial macro cells, particularly with regard to the possible RF radiation hazards.

Centralized architecture improves efficiency in resource utilization, i.e. traffic dimensioning can be sized according to the average traffic in the entire service area instead of the busy hour traffic since resources can be shared among all cells.

Inherent Synchronization among different cells due to the possibility of implementing a single timer, allowing Faster and softer intra-HAPS handover.

Increase in system capacity is possible through reduction of the cell size by antenna beam shaping.

Upgrading of the equipment can be easily done at a central location.

APPLICATIONS:

For high speed wireless communications
One of latest use of HAPs has been for wireless communications. Scientists are considering them as a platform to deliver high speed connectivity to users, over areas of up to 400 km.

For real-time monitoring of a region
Another future use which is currently being investigated is monitoring of a particular area or region for activities such as flood detection, seismic monitoring, and remote sensing as well as for disaster management.

For weather/ environmental monitoring and studying
Perhaps the most common use of high altitude platforms is for environment/ weather monitoring. Numerous experiments are conducted through high altitude balloon mounted with scientific equipment, which is used to measure environmental changes or to keep track of weather.

Military Use
While the commercial communications industry and the public safety community does not typically use deployable aerial communications architecture today, the U.S. military uses aerial platforms for signal transmission. The range extension and additional coverage area provided by aerial platforms are appealing in a tactical and dynamic battle space. The military has
employed aerial platforms using piloted aircraft, unmanned aerial vehicles (UAVs), and tethered or un tethered balloons for localized communications and to provide enhanced coverage areas and extend the battle space. The HAP technology has its own particular characteristics and capabilities, but all capitalize on the unique propagation advantages that altitude provides and use coordinated frequency assignments to allow multiple users on the ground to access the aerial platform and enjoy the increase in coverage area.[11]

CHALLENGES:

Lightweight strength of engine wind factor Getting them up and down safety

7.0 FUTURE:

HAPS will be deployed together with terrestrial and satellites elements to provide another degree of flexibility for system deployment that can be easily adjusted to the needs of the network operators and users' traffic demands.
HAPS will play a complementary role in future mobile system infrastructure e.g. consisting of W-LAN, cellular and satellite mobile systems to ease the deployment and roll out of the 3G and beyond 3G services

8.0 CONCLUSION:

We come to the conclusion that the HAP stations are a very beneficial and efficient approach in the advent of 5G technology in the near future.

9.0 REFERENCES:

[1] http://www.ieee.org

[2] http://www.skylink.aero/

[3] http://www.freewimaxinfo.com

[4] http://en.wikipedia.org

[5] http://wireless.ictp.it/school_2007/haps.pdf

[6] http://network.ee.tsinghua.edu.cn

[7] http://www.economist.com/node/2246155

[8] *http://en.wikipedia.org/wiki/X_band*

[9] Z. Peng and D. Grace, "Coexistence Performance of High-Altitude Platform and Terrestrial Systems Using Gigabit Communication Links to Serve Specialist Users", EURASIP Journal on Wireless Communications and Networking, Volume 2008, Article ID 892512, 11 pages doi:10.1155/2008/892512

[10] Zhe Yang and Abbas Mohammed, "High Altitude Platforms for wireless mobile communication applications", Blekinge Institute of Technology, Sweden

[11] White Paper: The Role of Deployable Aerial Communications Architecture in Emergency Communications and Recommended Next Steps

[12] Ha Yoon Song, "A Method of Mobile Base Station Placement for High Altitude Platform based Network with Geographical Clustering of Mobile Ground Nodes", Proceedings of the International Multi conference on Computer Science and Information Technology pp.869–876

IT 017

FEATURE BASED AUDIO SEGMENTATION USING k-NN AND GMM METHODS

Name of Main Author :Mrs.Borawake Madhuri Pravin

Designation : Lecturer

Name of organization :College of Engineering ,Manajri,41207

Pune,411028 Maharashtra, India

madhuri.borawake@gmail.com,madhuri_borawake@yahoo.co.in

Contact : 9823353507,9823353524

Research Scholar form JJT,University

Name of Co-Author :Prof..Kawitkar Rameshwar

Designation : Professor

Name of organization :Sinhgad College Of Engineering, Pune

Pune, Maharashtra, India

rskawitkar@rediffmail.com

Contact : 9890551983

Name of Co-Author :Prof..Khadtare Mahesh

Designation :Research Scholar

Name of organization :International Institute of Information Technology, Pune

Pune, Maharashtra, India

maheshkha@gmail.com

Contact : 9850187136

ABSTRACT : This project describes the work done on the development of an audio segmentation and classification system. Many existing works on audio classification deal with the problem of classifying known homogeneous audio segments. In this work, audio recordings are divided into acoustically similar regions and classified into basic audio types such as speech, music or silence. Audio features used in this project include Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate and Short Term Energy (STE). These features were extracted from audio files that were stored in a WAV format. Possible use of features, which are extracted directly from MPEG audio files, is also considered. Statistical based methods are used to segment and classify audio signals using these features. The classification methods used include the General Mixture Model (GMM) and the k- Nearest Neighbour (*k*-NN) algorithms. It is shown that the system implemented achieves an accuracy rate of more than 95% for discrete audio classification.

Keywords: *audio content analysis, segmentation, classification, GMM, k-NN, MFCC, ZCR, STE and MPEG*

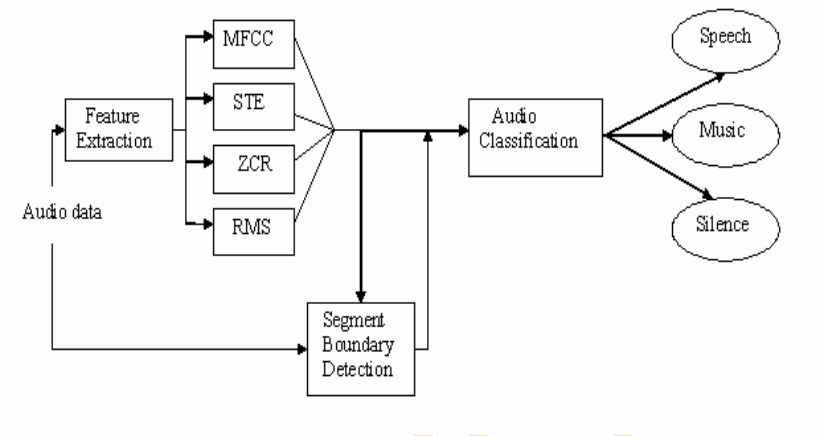
Introduction

Audio segmentation and classification have applications in wide areas. For instance, content based audio classification and retrieval is broadly used in the entertainment industry, audio archive management, commercial music usage, surveillance, etc. There are many digital audio databases on the World Wide Web nowadays; here audio segmentation and classification would be needed for audio searching and indexing. Recently, there has been a great deal of interest in monitoring broadcast news programs, in this case classification of speech data in terms of speaker could help in efficient navigation through broadcast news archives.

Like many other pattern classification tasks, audio classification is made up of two main sections: a signal processing section and a classification section. The signal processing part deals with the extraction of features from the audio signal. The various methods of time-frequency analysis developed for processing audio signals, in many cases originally developed for speech

processing, are used. The classification part deals with classifying data based on the statistical information extracted from the signals.

Two different classifiers, k-Nearest Neighbour(k-NN) and General Mixture model (GMM), were trained and tested to classify audio signals into music, speech and silence. The audio features used for classification were the Mel Frequency Cepestral Coefficients(MFCC), Zero Crossing Rates(ZCR) and Short Time Energy(STE). And for segmentation purposes Root Mean Square(RMS) features were used.



Segmentation and classification of audio data.

Audio feature extraction

Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying characteristic information about the signal. These vectors are used as basis for various types of audio analysis algorithms. It is typical for audio analysis algorithms to be based on features computed on a window basis. These window based features can be considered as short time description of the signal for that particular moment in time.

The performance of a set of features depends on the application. The design of descriptive features for a specific application is hence the main challenge in building audio classification systems. A wide range of audio features exist for classification tasks. These features can be divided into two

categories: time domain and frequency domain features. The Features considered in this chapter are: Mel Frequency Cepstral coefficient (MFCC), zero crossing rates and short time energy.

Audio Classification

The problem of classifying the extracted audio features into one of a number of audio classes is considered. The basic classification task can be considered as a process

where a previously unknown input data is assigned to a class

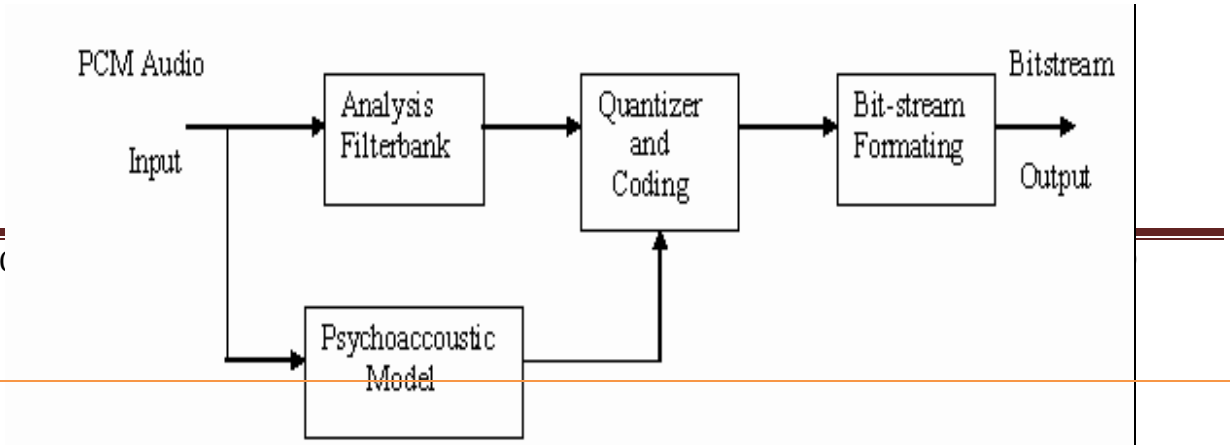
$C \in \{C_1, C_2, ..., C_n\}$. Such

assignments are made by establishing and applying a decision rule; for example, a simple decision rule could be the assignment of a new data sample to a class whose mean it is closest to in feature space.

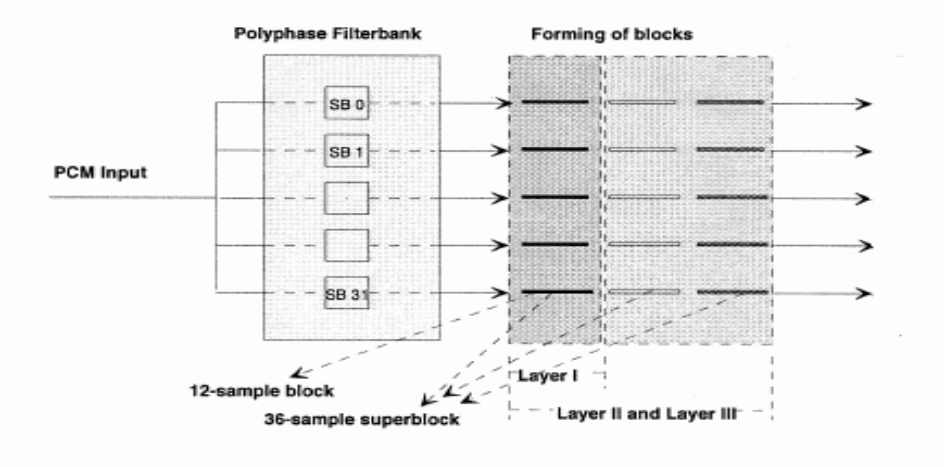
Classification algorithms are divided into supervised and unsupervised algorithms. In a supervised classification, a labelled set of training samples is used to "train" the algorithm whereas in the case of an unsupervised classification the data is grouped into some clusters without the use of labelled training set. Parametric and nonparametric classification is another way of categorizing classification algorithms. The functional form of the probably density of the feature vectors of each class is known in parametric methods. In non parametric methods on the other hand, no specific functional form is assumed in advance, instead the probability density is rather approximated locally based on the training data.

MPEG Audio Compression

In the following, a short description of the coding methods for the three MPEG-1 layers is given.



Block diagram of MPEG encoding



Subband blocks in MPEG encoding

Description of the audio data

The audio files used in the experiment were randomly collected from the internet and from the audio data base at IMM. The speech audio files were selected from both Danish and English language audios, and included both male and female speakers. The music audio samples were selected from various categories and consist of almost all musical genres. These files were in different formats (MP3, aif, wav, etc) and in order to have a common format for all the audio files and to be able to use them in matlab programs, it was necessary to convert these files to a wav format with a common sampling frequency. For this purpose the windows audio recorder was used and the recorded audio files were finally stored as 22050 Hz, 8 bit, mono audio files. The recorded audio files were further partitioned into two parts: the training set and the test set. This was important since each audio file was intended to be used only once, either for training or for testing a classifier. The training vectors correspond to 52566 frames for speech and 73831frames for music files.

Audio type	Number of files	Average length	Total length
Speech	45	15 sec.	675 sec.
Music	55	15 sec.	825 sec.

Training data

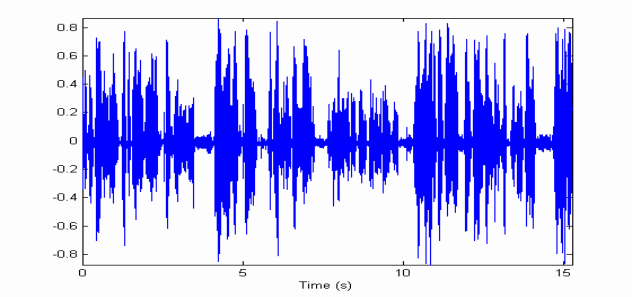
Audio type	Number of files	Average length	Total length
Speech	30	15 sec.	450 sec.
Music	30	15 sec.	450 sec.
silence	30	15 sec.	450 sec.
Music +	10	120 sec.	1200 sec.

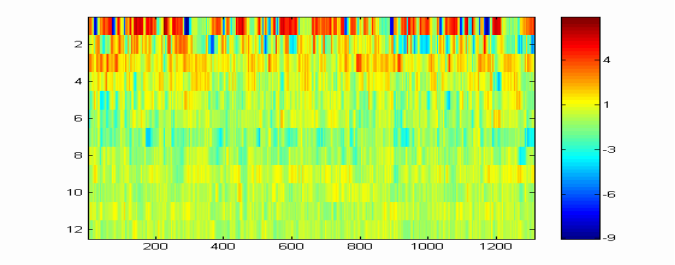
Test data

MFCC features

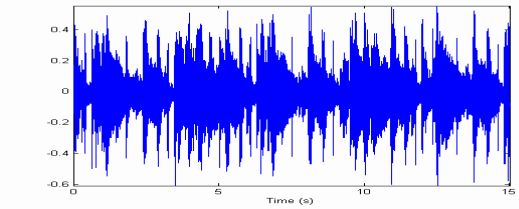
In order to extract MFCC features from the row audio signal, the signal was first partitioned into short overlapping frames each consisting of 512 samples. The overlap size was set to half the size of the frame. A Hamming window was then used to window each frame to avoid signal discontinuities at the beginning and end of each frame. A time series of MFCC vectors are then computed by iterating over the audio file resulting in thirteen coefficients per frame. The actual features used for classification task were the means of the MFCCs taken over a window containing 15 frames. Furthermore only six out of the thirteen coefficients were used. In this way a very compact data set was created. The following figures show plots of the speech and music signals as a function of time together with

their respective MFCCs .

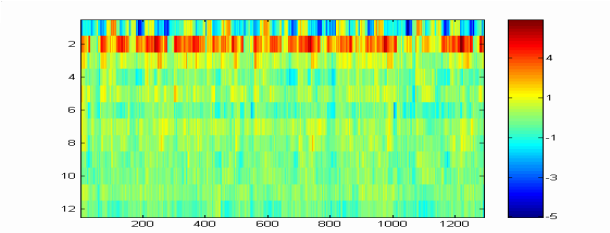




Plot of the MFCCs for the speech signal function of time



Plot of a speech signal as

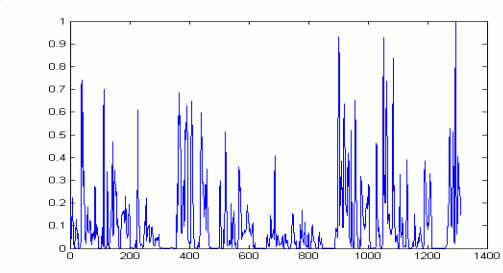


Plot of a music signal as function of time

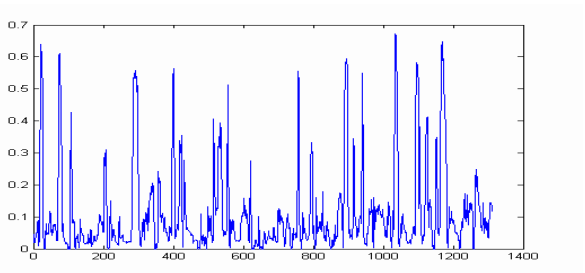
Plot of the MFCCs for the music signal

The STE and ZCR features

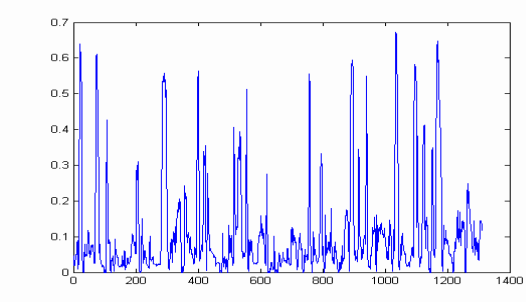
Since these features were intended to be used either in conjunction with the MFCCs or independently, it was necessary to split the audio signal so that the length of these features were the same as the length of the MFCCs. Hence, the partition of the audio signal into overlapping windows was exactly the same as in the case of the MFCC features. The Short Time Energies and the Zero-crossing rates were extracted from such windows, one from each window. The actual features used for the classification task were the means taken over a window containing 15 frames. The following figures show plots of STE and ZCR for both music and speech signals.



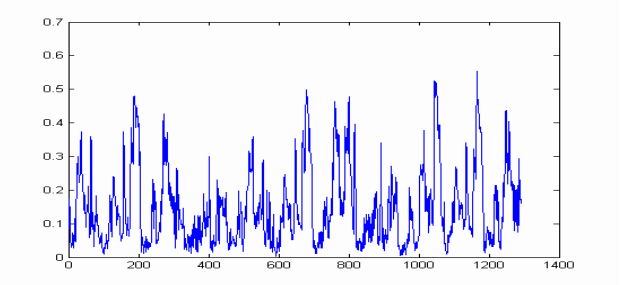
STE for speech signal



STE for music signal



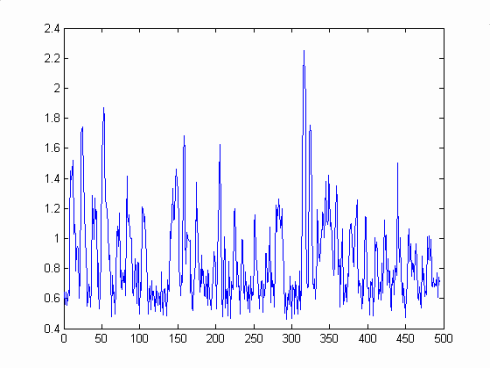
ZCR for speech signal



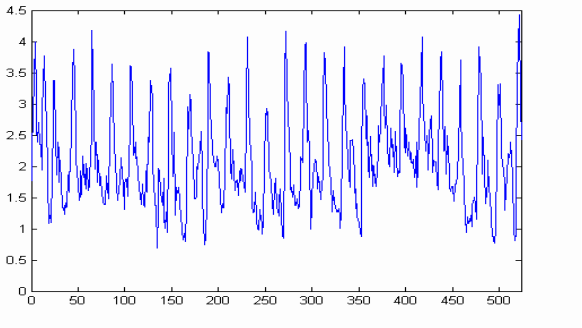
ZCR for music signal

The RMS feature

Although the RMS is somewhat related to the short time energy, it is often used as a measure of the loudness of audio signals and therefore a unique feature to segmentation. Since this feature was used alone for another task, the audio signal was split in a rather different way. The audio signal was first partitioned into short non overlapping frames each consisting of 512 samples. The Root Mean Square was computed by iterating over the audio file based on the amplitude equation shown on page 25 and a single RMS value is obtained for each frame. The following figures show plots of RMS for both music and speech signals.



RMS of a speech signal



RMS of a music signal

	Music	**Speech**
Music	**89.1**	10.9

Speech	8.4	**91.6**

Confusion matrix when MFCC features were the only inputs.

	Music	**Speech**
Music	**93.01**	6.99
Speech	5.49	**94.51**

Confusion matrix when the inputs were MFCC and STE features.

	Music	**Speech**
Music	**90.6**	9.39
Speech	7.4	**92.53**

Confusion matrix when the inputs were the MFCC and ZCR.

	Music	**Speech**
Music	**93.6**	6.33
Speech	5.7	**94.29**

Confusion matrix when the inputs were MFCC, ZCR and STE

features.

From the results obtained the following observations can be made. The MFCC features used as an input, alone, result in an overall correct classification rate of 90.3%. When the MFCC features were used in conjunction with the short time energy and the Zero Crossing rate the overall classification rate gets better and is around 93.98%. The same is true when MFCC feature are used together with short time energy features. However, when the input to the classifier was a combination of MFCC features and zero crossing

rate only little improvement in the overall correct classification rate was seen. We conclude therefore that the MFCC features in conjunction with the short time energy alone can with a good classification rate be used for a speech/music discrimination.

	Music	**Speech**
Music	**85.22**	**14.78**
Speech	**0.4**	**99.56**

The features used were the MFCCs.

	Music	**Speech**
Music	**89.78**	**10.22**
Speech	**0.22**	**99.78**

The features used were the MFCC and STE features.

	Music	Speech
Music	85.65	14.35
Speech	0.00	100.00

The features used were the MFCC and ZCR features.

	Music	Speech
Music	91.30	8.70
Speech	0.00	100.00

The features used were the MFCC, STE and ZCR features.

Although the results obtained in this case showed similar tendencies as in the case of the K-nearest neighbour classifier, the correct classification rate was even better. When the MFCC features were used in conjunction with the short time energy and zero crossing rate, a correct classification rate of around 95.65% was obtained. This result was the best result among the classification results obtained from both the GMM classifier and the KNN classifiers. A correct classification rate of about 94.78% was obtained for the case when MFCC in conjunction with the Short Time Energy features were used. However, for the case where the input was a combination of MFCC and ZCR features, the classification rate was 92.83% , which is almost the same as when pure MFCC features were used.

Comparison of the classification results

The Table below shows the classification results obtained for the two classifiers with different feature combinations.

Features	**k-NN (k=5) Accuracy (%)**	**GMM Accuracy (%)**
MFCC	90.35	92.39
MFCC + ZCR	91.57	92.83
MFCC + STE	93.76	94.78
MFCC + ZCR+ STE	93.98	95.65

Accuracy testing results for the speech/music classifier

The classification results obtained from using a general mixture model classifier and a k- nearest neighbour classifier demonstrate the effect of the classification algorithms. The general mixture model classifier seemed to have a far better correct classification rate than k-nearest neighbour classifier (around 2%). In both cases, adding more features to the MFFC features showed a positive effect on the outcome, although using the MFCC in conjunction with STE resulted in a relatively higher classification rate than when MFCC features were used in conjunction with the zero crossing rates.

Conclusion and future work

The aim of this project was to design a system that could be used to segment an audio signal into similar regions and then classify these regions into music, speech and silence audio classes. The project could be considered as a combination of two tasks; a segmentation task and a classification task. Classification algorithms were used either independently with a given audio segment or in combination with the segmentation algorithm.

Features extracted from music and speech signals (in WAV format) were used in the two tasks. Three feature sets were used to train and test two different classifiers, the General Mixture Model classifier and the k-Nearest Neighbour classifiers, to classify audio signals, and only one feature set was used to partition audio into similar regions. Nearly all the audio files used in this

project had been obtained from the internet. The majority of these audio files were in MP3 format and it was necessary to convert them to WAV format. Thus, the process for extracting audio feature showed to be very time consuming. It would have been very advantageous if the system was designed to take in audio in MP3 format. This could have had two effects on the system; the need for converting one audio format to another would have been avoided, and features would have been extracted directly from the encoded data. The two classifiers were trained and tested with the same training and test sets. With each classifier, four experiments were run with different combinations of the feature sets.

The system implemented worked well on classifying any type of music and speech segments with a correct classification rate of 95.65% for one second windows. The system also worked reasonably well for segmenting audio signals into similar classes. Some improvement in the segmentation method used is however required.

There are many things that could be done in the future. The segmentation algorithm could be modified to detect the transition point with an accuracy of 30ms, and also to automatically set the threshold for finding the local maxima of the normalised distance measure. More training data could be used in the classification part. The system could be trained to include other classes other than music, speech and silence. Further classifications into different music genre or identifying a speaker are also other possibilities.

References

[1] Lie Lu, Hong-Jiang Zhang and Hao Jiang. "Content analysis for audio classification and segmentation". *IEEE Transactions on speech and audio processing,* vol.10, no.7, October 2002

[2] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal , " Speech/Music discrimination for multimedia applications," *Proc. IEEE Int. Conf. on acoustics, Speech, Signal Processing* (Istanbul), pp. 2445-2448, June 2000

[3] H. Meindo and J.Neto, " Audio Segmentaion, Classification and Clustering in a Broadcast News Task" , *in Proceedings ICASSP 2003*, Hong Kong, China, 2003.

[4] G. Tzanetakis and P. Cook, " Multifeature audio segmentation for browsing and annotation," *Proc.1999 IEEE workshop on applications of*

signal processing to Audio and Acoustics, New Paltz, New York, Oct17-20, 1999.

[5] C. Panagiotakis and G.Tziritas " A Speech/Music Discriminator Based on RMS and Zero-Crossings". *IEEE Transactions on multimedia*, 2004.

[6] E. Scheirer and M. Slaney, " Construction and evaluation of a robust multifeature speech/music discriminator, " in *Proc. ICASSP '97*, Munich, Germany, 1997, ,. 1331-1334.

[7] Davis Pan, "A Tutorial on MPEG/Audio Compression,". IEEE Multimedia Vol. 2, No.7, 1995, pp. 60-74.

[8] Silvia Pfeiffer and Thomas Vincent "Formalisation of MPEG-1 compressed domain audio features", Technical Report No.01/196, CSIRO Mathematical and Information Sciences, Dec. 2001.

[9] G. Tzanetakis and P. Cook, " Sound analysis using MPEG compressed audio", *Proc. IEEE Intl. Conf. on acoustics, Speech, Signal Processing,* ICASSP, 2000

[10] D. Pan, " A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, No.2, 1995, pp.60-74.

[11] Christopher M. Bishop, *Neural Networks for Pattern Recognition* , Oxford University Press, 1995

[12] Tong Zhang and C.C. Jay Kuo, "Heuristic Approach for Generic Audio Data Segmentation and Annotation," *ACM Multimedia (1),* 1999, pp 67-76.

[13] Beth Logan, " Mel Frequency Cepstral Coeffcients for Music Modelling," *in international Symposium on Music information retrieval*, October 2000.

[14] John R. Deller, Jr., John H.L. Hansen and John G. Proakis, *Discrete-Time Processing of Speech Signals,* IEEE Inc. 2000.

[15] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing principles, algorithms and applications*, Prentice-Hall, Inc, 1996.

[16] L.R. Rabiner and R.W.Schafer, *Digital Processing of speech signals*, Prentice-Hall, 1978.

[17] MPEG Maaate. http://www.cmis.csiro.au/Maaate/

IT 018

Governance of IT in Indian Banks – An Immense Need

Mrs. K. S. Pawar[1], Dr. R. D. Kumbhar[2]

Asst. Prof1[1], College of Computer Application for Women, Satara,
pawar.kishori1@gmail.com

HOD[2], Computer Dept., KBPIMSR, Satara, rdk14@rediffmail.com

Abstract:

Information technology has become an essential component of any financial organization in the present global scenario. Information Technology governance is a concept that has suddenly emerged and become an important issue in the information technology field. It has been observed by researcher banks in India are using IT based systems from last three decades for performing different banking functions but their implementation is done in a disorganized manner. Therefore implementation of IT is not as effective as it should be. In fact implementation is not everything all but its prefect governance is a must to get the optimum benefits of IT enabled modern technologies. For getting real fruits of IT, proper IT governance is required. This article tries to highlight importance of IT governance, obstacles in IT governance and implementation framework for IT governance in banks.

Keywords: IT Governance, BCG, ITGI, COBIT

1. Introduction:

Banking in India originated in the last decades of the 18th century. The oldest bank in existence in India is the State Bank of India, which originated in the Bank of Calcutta in June 1806. The Indian banking system comprises the following institutions:

1. Commercial banks

    a. Public sector

    b. Private sector

    c. Foreign banks

    d. Cooperative institutions

        (i) Urban cooperative banks

        (ii) State cooperative banks

        (iii) District Central Cooperative banks

The Indian banks were finding it difficult to compete with the international banks in terms of the customer service without the use of the information technology and computers. The IT revolution had a great impact in the Indian banking system. The use of computers had led to introduction of online banking in India. The use of the computerization in banking sector of India has increased after the economic liberalization of 1991 as the country's banking sector has been exposed to the world's market.

Today's business environment is very dynamic and undergoes rapid changes as a result of technological innovation, increased awareness and demands from customers. The banking industry of the 21st century operates in a complex and competitive environment characterized by these changing conditions and highly unpredictable economic climate. Information and Communication Technology is at the centre of this global change curve. The application of IT concepts and techniques has become a subject of fundamental importance to all banks. Because of these all reasons banks have implemented IT technologies but their implementation is being done in a disorganized manner. Therefore implementation of IT is not as effective as it should be. To get real fruits of IT, proper IT governance is required. IT governance is essential to mitigate IT related risks and avoid IT project failures.

2. Present Status of IT in Banking:

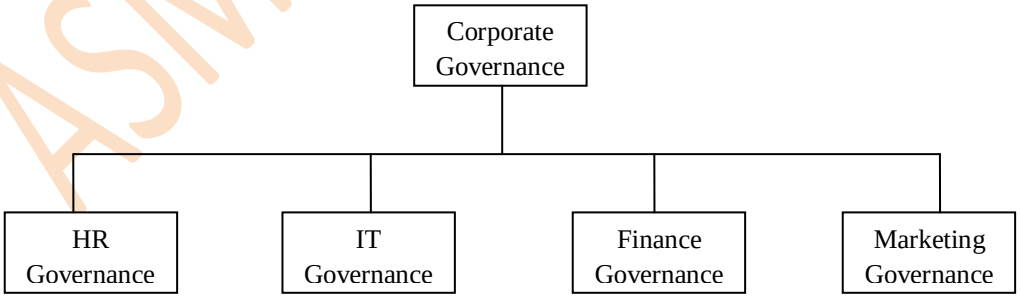
Tremendous improvement has taken place in the Indian banking sector because of the IT revolution. Most of the private and foreign banks have gone for 100 percent computerization. At a rapid fast, the government owned nationalized banks are also improving their number of computerized branches. With the help of computerization, the transaction takes place at a faster rate and the waiting time of a customer in a queue is getting minimized. E – Banking is emerging in the Indian banking sector. The banks provide the facility of internet banking, phone and mobile banking with the help of IT. All the branches of a bank are networked. The networking also takes place between two or more branches in order to provide easy accessibility for a customer. The cost incurred on infrastructure, furniture and employees has got decreased because of the application of IT.

3. IT Governance:

IT governance is the responsibility of the board of directors and executive management. It is an integral part of corporate governance and consists of the leadership, organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategies and objectives.

Weill and Ross define IT governance as "Specifying the decision rights and accountability framework to encourage desirable behavior in the use of IT."

Following figure shows where IT governance fIT 0



IT governance is a subset discipline of corporate governance focused on information technology (IT) systems and their performance and risk management. It is indicted that IT management is also involved in the

governance process. However, a clear difference must be made between IT management and IT governance. IT management focuses on the daily effective and efficient supply of IT services and operations. IT governance, in turn, is much broader and concentrates on performing and transforming IT to meet present and future demands of the business and customers. IT Governance focuses specifically on information technology systems, their performance and risk management. The primary goals of IT Governance are to assure that the investments in IT generate business value, and to mitigate the risks that are associated with IT. This can be done by implementing an organizational structure with well-defined roles for the responsibility of information business processes, applications and IT infrastructure. While designing IT governance, it is important to recognize that it is depend upon a variety of internal and external factors. Therefore, selecting the right mechanism is a complex process and what works for one organization may not necessarily work for another, even if they work in the same sector.

Need of IT Governance:

At present banks are implementing IT applications in different areas but these banks are not getting real benefits due to improper governance of IT. Following are the obstacles observed in banks IT governance.

Top management does not emphasize IT

Shrinking the responsibility by the concerned authorities

Poor strategic alignment

Ineffective resource management

IT staffing problem

No review on IT performance

Security & privacy incidents

Lack of training.

Banking environment has become highly competitive today. To be able to survive and grow in the changing market environment banks are going for the latest technologies. IT has become a major enabler to almost all business transformation initiatives. How IT is being used will have a very important impact on whether the organization achieves its vision, mission or strategic goals. Information Technology has also provided banking industry with the ability to deal with the challenges the new economy creates. IT directly affects how managers decide, how they plan and what products and services are offered in the banking industry. It has continued to change the way banks and their corporate relationships are organized worldwide and the variety of innovative devices available to enhance the speed and quality of service delivery.

IT is now so fundamental and persistent within enterprises that governance needs to pay special attention to IT, reviewing how strongly the banks relies on IT and how critical IT is for the execution of the business strategy because IT is critical in supporting and enabling enterprise goals and IT is strategic to the business growth and innovation.

Business process transformation is very difficult without adequate IT governance. IT governance is essential to mitigate IT related risks and avoid project failure. Ineffective IT governance is likely to be a root cause of the negative results.

Following are three main reasons for importance of IT Governance -

a) Value – IT is typically an expensive asset in any type of organization.

b) Risk – Organizations are become dependent on IT.

c) Alignment - Overall strategy is very much dependent upon the IT strategy.

4. Areas of IT Implementation in Banks:

The software packages for banking applications in India had their beginnings in the middle of 80s, when the banks started computerizing the branches in a limited manner. The early 90s saw the reducing hardware prices and advent of cheap and inexpensive but high powered PC's and services and banks went in for what was called Total Branch Automation (TBA) packages. In the middle and late 90s there was revolution in communication technologies like internet, mobile/cell phones etc. Technology has continuously played important role in the working of banking institutions and the services provided by them like safekeeping of public money, transfer of money, issuing drafts, exploring investment opportunities and lending drafts, exploring investment being provided. IT is increasingly moving from a back office function to a prime assistant in increasing the value of a bank over time.

Following are areas of IT implementation in banks –

a) Deposits and Advances

b) Investments

c) Services offered to customers

Credit cards/ Debit cards

ATM

E-Cheques

EFT (Electronic Funds Transfer)

DeMAT Accounts

Mobile Banking

Telephone Banking

Internet Banking

EDI (Electronic Data Interchange)

Bill Payment

Smart Money Order

Online Payment of Excise & Service Tax

d) Human Resource Management

Manpower planning, recruitment & placement

Attendance & compensation management

Personnel information system

Training & development

e) Financial Accounting

5. IT Governance Framework for Banks

Fundamentally, IT governance is concerned with two things: Value delivery of IT to the business and mitigation of IT risks. The first is driven by strategic alignment of IT with the business. The second is driven by embedding accountability into the bank. Both need to be supported by adequate resources and measured to ensure that the results are obtained. This leads to the five main focus areas for IT governance. Two of them are outcomes: value delivery and risk

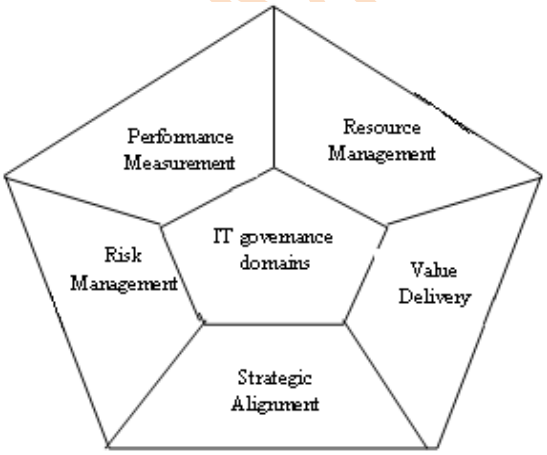
management. Three of them are drivers: strategic alignment, resource management and performance measurement.

Each bank operates in and environment that is influenced by

- Stakeholders (stakeholders are customers, business organizations, shareholders, insurance companies, RBI, other apex financial institutions)
- The community
- Policies, Rules and Regulations of RBI and applicable laws.
- Banking practices

Focus Areas for IT Governance in Banks:

One of the well-known international frameworks in achieving effective control over IT and related risks is the "Control Objectives for Information Technology" (COBIT) that is issued by ITGI. The framework provides five focus areas for IT Governance and shown as below.



- IT strategic alignment
- Value delivery
- IT resource management
- IT risk management
- Performance measurement

a) IT strategic alignment

Strategic alignment deals with the key question–whether a bank's technology investment is aligned to its strategic business objectives, enabling the formation of capabilities necessary to deliver business value. IT strategy provides banks the opportunity to:

- Reduce costs and improve administrative efficiency
- Increase managerial effectiveness
- Add value to products and services
- Assist in competitive positioning

While formulating an IT strategy, a bank must consider:

- Operating cost of current IT: whether this provides sufficient value to the business
- Business objectives and competitive environment
- Current and future technologies: costs, risks and benefits

- Capability of the IT organization and technology to deliver current and future levels of service and its implication on the bank (degree of change and investment)

As IT gets more critical for a bank's survival in addition to enabling growth, IT Strategy Committees need to broaden their scope beyond offering advice on strategy, to other areas like IT risks, value and performance.

Challenges in IT Strategy:

- Ensuring business and IT goals cascade throughout the bank into roles, responsibilities and actions
- Ensuring an effective communication and engagement between business and IT management
- Identifying barriers to strategic alignment
- Evaluating effectiveness of alignment of IT and strategic business initiatives
- Monitoring and assessing current and future technology improvements

With Respect to IT Strategic Alignment, Banks need to ensure the following:

i)Major IT development projects need to be aligned with business strategy

ii) Banks should have up-to-date business strategy that sets out a clear direction for IT that is in accordance with the business objectives

iii)IT investments need to be suitably balanced between maintaining the infrastructure that support the bank's "as is" operations, and the infrastructure that transforms the operations and enables the business to grow and compete in new areas

iv) IT budget reflects priorities established by the IT 0related investment programmes and includes ongoing costs of maintaining the infrastructure

v)Board's IT Strategy Committee reviews and advises the management about IT 0related investments

vi) IT Steering Committee (or equivalent) composed of executives from business and IT management has responsibility to: determining prioritization of IT 0related investment, track status of projects, resolve resource conflict, monitor service levels and service improvements

vii) Performance of IT management is monitored

b) Value delivery

The basic principles of IT value delivery are on time and within budget delivery of IT projects, with appropriate quality, which achieves benefits that were promised. For effective IT value delivery to be achieved, both actual costs and Return on Investment (ROI) need to be managed.

The bank should set expectations relative to IT deliverables:

- Fit for purpose and meeting business requirements
- Flexibility to adopt future requirements
- Throughput and response times
- Ease of use and security
- Integrity, accuracy and confidentiality of information

Alignment of technology to business provides value by delivering infrastructure that enable the bank to grow by improving customer satisfaction, assuring

customer retention, breaking into new markets, increasing overall revenue and driving competitive strategies.

With respect to "value delivery", banks need to ensure that:

i) IT enabled investment programmes and other IT assets and services are managed to determine that they deliver the greatest possible value in supporting the bank's strategy and objectives

ii) Effective IT controls are place to minimize IT related risks, increase efficiency, use resources optimally and increase the effectiveness of IT processes

iii) IT function must supports Management Information System in respect of decision making by management

iv) Project management and quality assurance steps should be implemented to ensure systems are delivered on time, to cost and with the necessary level of functionality

v) IT internal control failures and weaknesses and their actual and potential impact need to be evaluated and management takes suitable actions in respect of such control failures or weaknesses

vi) Project-level steering committees needs to be created for taking responsibility for execution of the project plan, achievement of outcomes and project completion. The various responsibilities include reviewing progress against the project plan, reviewing and approving changes to project resource allocation, time lines, objectives, costs, keeping the project scope under control

vii) Independent assurance on the achievement of IT objectives and the control of IT risks is conducted regularly

viii) IT Steering Committee or any of its sub committees involving the CIO and senior business managers prioritize IT initiatives and assign ownership for IT 0enabled business opportunities

ix) Periodical review of all non-performing or irrelevant IT projects in the bank, if any, and taking suitable actions

c) IT Resource Management

A key to successful IT performance is optimal investment, use and allocation of IT resources: people, applications, technology, facilities and data, in servicing the bank's needs. Additionally, the biggest challenge has been to know where and how to outsource, and then to know how to manage the outsourced services in a way that delivers the values promised at an acceptable price.

IT assets are complex to manage and continually change due to the nature of technology and changing business requirements. Effective management of hardware life-cycles, software licenses, service contracts and permanent and contracted human resources is a critical success factor. It is critical not only for optimizing the IT cost base, but also for managing changes, minimizing service incidents and assuring a reliable service quality.

For IT resource management, banks should consider the following:

i) That the Board is aware of IT resources and infrastructure to meet strategic business objectives

ii) Policies and procedures for information systems monitoring facilitate, consistent and effective reporting and review of logging, monitoring and reporting of system events

iii) Information on IT investments is available to the Board and Senior Management

iv) Responsibilities, relationships, authorities and performance criteria of project team members and stakeholders are stated

v) Requirement for trained resources, with the requisite skill sets for the IT function, is understood and assessed. A periodic assessment of the training requirements for human resources is made to ensure that sufficient, competent and capable human resources are available

d) IT Risk Management

Effective risk management begins with identifying high-level risk exposures. Dependent on the type of risk, project and its significance to the business, Board and Senior Management may choose to take up any of the three actions:

- Mitigate—Implement controls (e.g. acquire and deploy security technology to protect the IT infrastructure)
- Transfer—Share risk with partners or transfer to insurance coverage
- Accept—Formally acknowledge that the risk exists and monitor it

At a basic level, risk should at least be analyzed, even if there is no immediate action to be taken, the awareness of risk will influence strategic decisions. An IT control framework defines stakeholders and relevant controls for effective Enterprise Risk Management. The “risk register”, usually in form of a table, is a tool that assists in risk management. It is also called a “risk log”. It usually is used when planning for the future that includes project, organizational, or financial plans. Risk management uses risk registers to identify, analyze and manage risks in a clear and concise manner. Risk register contains information on each identified risk and planned responses are recorded in the event the risk materializes, as well as a summary of what actions should be taken before hand to reduce the impact. Risks are ranked in order of likelihood, or of their impact and record the analysis and evaluation of risks that have been identified. The register or the log may be created for a new project or investment.

In respect to IT risk management, banks should consider the following:

i. IT management needs to assess IT risks and suitably mitigate them

ii. Bank-wide risk management policy, in which operational risk policy includes IT 0related risks, is in place. The Risk Management Committee periodically reviews and updates the same (at least annually)

iii. Bank's risk management processes for its e-banking activities are integrated into its overall risk management approach

iv. All risks related to suppliers are considered.

v. Operational risk inherent in all material products, activities, processes and systems, are assessed and relevant controls are implemented and monitored

vi. Information security policy is in place and requirements indicated in the chapter on information security are considered

vii. Comprehensive and centralized change control system is implemented at levels (project or application), so that changes are appropriately reviewed and approved

viii. For managing project risks, a consistent and formally-defined programme and project management approach needs to be applied to IT projects that enable stakeholder participation and monitoring of project risks and progress.

ix. Inter-dependencies between risk elements are considered in the risk assessment process

e) Performance Measurement

IT performance management aims at:

- Identifying and quantifying IT costs and benefits

- Overcoming limitations of traditional quantifiable performance measures (financial terms) such as ROI, Net Present Value (NPV), Internal Rate of Return (IRR) and payback method
- Overcoming limitations of measuring "unquantifiable" values

In respect to the IT performance management, the considerations for a bank are the following:

- That information on IT projects that have an impact on the bank's risk profile and strategy are reported to appropriate levels of management and undergo appropriate strategic and cost and reward analysis on a periodic basis
- Processes for making return versus risk balance may be considered and supported with standard templates or tools
- Tools such as IT balanced scorecard is considered for implementation, with approval from key stakeholders, to measure performance along dimensions: financial, customer satisfaction, process effectiveness, future capability and assess IT management performance based on metrics such as scheduled uptime, service levels, transaction throughput and response times and application availability
- The bank may also consider assessing the maturity level, set a target as per the IT Governance maturity model, design an action plan and subsequently implement it to reach the target maturity level
- Periodic assessment of IT budget deviations
- Periodic review and update of IS Policies and guidelines

Conclusion

Banking organizations initiated computerization in the middle of 80s. However, most of the areas covered under computerization are at operational level. This has resulted in poor information management, which ultimately resulted in poor and irrelevant decisions at the top level leading to ineffective governance. IT

governance is a very critical process; it needs to be implemented with right spirit with high level of commitment from top management and stakeholders of banks.

Adopting IT governance framework bank will create the foundation for improved business efficiency, decision making and resulting into good governance. Therefore proposed IT governance will enable banking sector for achieving greater heights both horizontally and vertically.

References

- Sanjay Anand, "Essentials of Corporate Governance (Essentials Series)", September 2007
- Wim Van Grembergen (Antwerp Management School, Belgium), "Strategies for Information Technology Governance", July 2003
- Koen Brand, Harry Boonen, Van Haren "IT Governance Based on Cobit 4.1: A Management Guide", 2008
- Devid Tetlock, "IT Governance Framework for Performance and Compliance", 2007
- http://www.docstoc.com/docs/15875696/Role-of-IT 0in-banking
- http://www.americanbanker.com/btn/17_11/-234286-1.html
- http://www.cio.com/article/111700/IT_Governance_Definition_and_Solutions#what
- http://www.highbeam.com/doc/1G1-177943283.html
- http://www.rbi.org.in/scripts/PublicationReportDetails.aspx?UrlPage=&ID=616
- http:// www.mbaknl.com/business-finance/role-of-information-technology-IT 0in-the-banking-sector
- http://www.bis.org/publ/bcbs56.html

IT 019

Tangible User Interface:Unifying The New Genration Of Interaction Styles

Yogesh H.Raje
Master of Computer Application
Institute of Information Technology, VIIT, Baramati,
MS, India.
yhraje@gmail.com

Dr.Amol C. Goje
Director
lnstitute of Information Technology ,VIIT, Baramati,
MS, India
amol@ict4rd.org

Abstract— Tangible User Interface (TUI) is a user interface in which a person interacts with digital information through the physical environment. Interactions with digital information are now largely confined to Graphical User Interfaces (GUIs). We are surrounded by a variety of ubiquitous GUI devices such as personal computers, handheld computers, and cellular phones. Tangible User Interfaces (TUIs) aim to take advantage of these haptic interaction skills, which is significantly different approach from GUI. The key idea of TUIs is to give physical forms to digital information. The physical forms serve as both representations and controls for their digital counterparts. TUI makes digital information directly manipulatable with our hands and perceptible through our peripheral senses by physically embodying it. This paper describes the basic concepts of TUI in comparison with GUI and genres of TUI applications to address the key properties of TUI and design challenges.

Keywords- TUI , GUI , Urp, Haptic

Introduction

Clearly People have developed sophisticated skills for sensing and manipulating their physical environments. However, most of these skills are not employed in interaction with the digital world today. A Tangible User Interface (TUI) is built upon those skills and situates the physically-embodied digital information in a physical space. Its design challenge is a seamless extension of the physical affordance of the objects into digital domain [1][2].

Interactions with digital information are now largely confined to Graphical User Interfaces (GUIs). We are surrounded by a variety of ubiquitous GUI devices

such as personal computers, handheld computers, and cellular phones. The Graphical User Interface (GUI) has been in existence since the 70's and the first appeared commercially in the Xerox 8010 Star System in 1981 [3]. With the commercial success of the Apple Macintosh and Microsoft Windows, the GUI has become the standard paradigm for Human Computer Interaction (HCI) today.

GUIs represent information (bits) with pixels on a bIT 0mapped display. Those graphical representations can be manipulated with generic remote controllers such as mice and keyboards. By decoupling representation (pixels) from control (input devices) in this way, GUIs provide the malleability to emulate a variety of media graphically. By utilizing graphical representation and "see, point and click" interaction, the GUI made a significant improvement over its predecessor, the CUI (Command User Interface) which required the user to "remember and type" characters.

However, interactions with pixels on these GUI screens are inconsistent with our interactions with the rest of the physical environment within which we live. The GUI, tied down as it is to the screen, windows, mouse and keyboard, is utterly divorced from the way interaction takes place in the physical world. When we interact with the GUI world, we can't take advantage of our dexterity or utilize our skills for manipulating various physical objects such as manipulation of building blocks or the ability to shape models out of clay.

Tangible User Interfaces (TUIs) aim to take advantage of these haptic interaction skills, which is significantly different approach from GUI. The key idea of TUIs is to give physical forms to digital information. The physical forms serve as both representations and controls for their digital counterparts. TUI makes digital information directly manipulatable with our hands, and perceptible through our peripheral senses by physically embodying it.

Tangible User Interface serves as a special purpose interface for a specific application using explicit physical forms, while GUI serves as a general purpose interface by emulating various tools using pixels on a screen

This paper describes the basic concept of TUI in comparison with GUI, early prototypes of TUI that highlights the basic design principles, and discusses design challenges that TUI needs to overcome.

Urp:an example of tui

To illustrate basic TUI concepts, we introduce "Urp" (Urban Planning Workbench) as an example of TUI

(Underkoffler and Ishii, 1999). Urp uses scaled physical models of architectural buildings to configure and control an underlying urban simulation of shadow, light reflection, wind flow, etc. (Photo1). In addition to a set of building models, Urp also provides a variety of interactive tools for querying and controlling the parameters of the urban simulation. These tools include a clock tool to change a position of sun, a material wand to change the building surface between bricks and glass (with light reflection), a wind tool to change the wind direction, and an anemometer to measure wind speed.

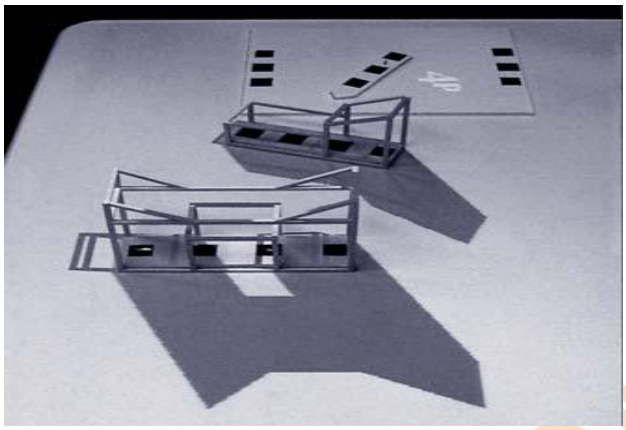


Photo 1 Urp and shadow simulation

Physical building models casting digital shadows, and a clock tool to control time of the day (position of the sun).

The physical building models in Urp cast digital shadows onto the workbench surface (via video projection),corresponding to solar shadows at a particular time of day. The time of day, representing the position of the sun,can be controlled by turning the physical hands of a "clock tool" (Photo 2). The building models can be moved and rotated, with the angle of their corresponding shadows transforming according to their position and time of day.

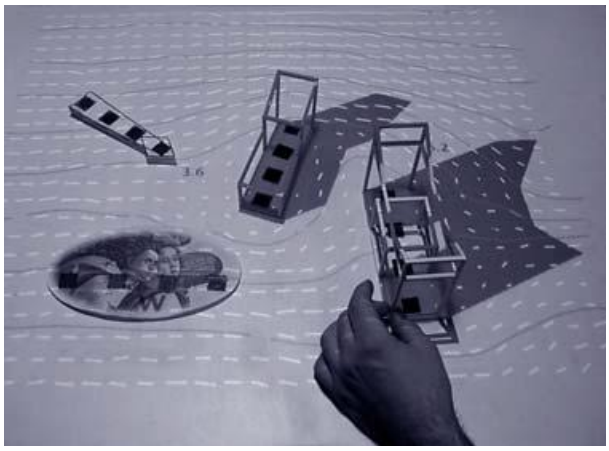


Photo 2 Urp and wind simulation

Wind flow simulation with a wind tool and an anemometer.

Correspondingly, moving the hands of the clock tool can cause Urp to simulate a day of shadow movement between the situated buildings. Urban planners can identify and isolate inter-shadowing problems (shadows cast on adjacent buildings), and reposition buildings to avoid areas that are needlessly dark areas, or maximize light between buildings.

A "material wand" alters the material surface properties of a building model. By touching the material wand to a building model, the building surface material is switched from bricks to glass, and a projected reflection of sunlight appears to bounce off the walls of the building. Moving the building allows urban designers to be aware of the relationship between the building reflection and other infrastructure. For example, the reflection off the building at sundown might result in distraction to drivers on a nearby highway. The designer can then experiment with altering the angles of the building to oncoming traffic or move the building further away from the roadway. Tapping again with the material wand changes the material back to brick, and the sunlight reflection disappears, leaving only the projected shadow.

In "Urp," physical models of buildings are used as tangible representations of digital models of the buildings. To change the location and orientation of buildings, users simply grab and move the physical model as opposed to pointing and dragging a graphical representation on a screen with a mouse. The physical forms of Urp's building models, and the information associated with

their position and orientation upon the workbench represent and control the state of the urban simulation. Although standard interface devices for GUIs such as keyboards, mice, and screens are also physical in form, the role of the physical representation in TUI provides an important distinction. The physical embodiment of the buildings to represent the computation involving building dimensions and location allows a tight coupling of control of the object and manipulation of its parameters in the underlying digital simulation.

In the next section, the model of TUI is introduced in comparison with GUI to illustrate this mechanism.

basic model of tangible user interface

The interface between people and digital information requires two key components; input and output, or control and representation. Controls enable users to manipulate the information, while external representations are perceived with the human senses.

GUI

In 1981, the Xerox Star workstation set the stage for the first generation of GUI [4], establishing the "desktop metaphor" which simulates a desktop on a bIT 0mapped screen. The Star workstation was the first commercial system that demonstrated the power of a mouse, windows, icons, property sheets, and modeless interaction. The Star also set several important HCI design principles, such as "seeing and pointing vs. remembering and typing," and "what you see is what you get (WYSIWYG)." The Apple Macintosh brought this new style of HCI to the public's attention in 1984, creating a new trend in the personal computer industry. Now, the GUI is widespread, largely through the pervasiveness of Microsoft Windows, PDAs, and cellular phones.

GUI uses windows, icons, and menus made of pixels on bitmapped displays to visualize information. This is an intangible representation. GUI pixels are made interactive through general "remote controllers" such as mice, tablets, or keyboards. In the pursuit of generality, GUI introduced a deep separation between the digital (intangible) representation provided by the bIT 0mapped display, and the controls provided by the mouse and keyboard.

TUI

Tangible User Interface aims at a different direction from GUI by using tangible representations of information which also serve as the direct control mechanism of the digital information. By representing information in both tangible and intangible forms, users can more directly control the underlying digital representation using their hands.

The tangible representation helps bridge the boundary between the physical and physical worlds. The tangible representation is computationally coupled to the control to the underlying digital information and computational models. Urp illustrates examples of such couplings, including the binding of graphical geometries (digital data) to the physical building models, and computational simulations (operations) to the physical wind tool.

Instead of using a GUI mouse to change the location and angle graphical representation of a building model by pointing, selecting handles and keying in control parameters, an Urp user can grab and move the building model to change both location and angle.

The tangible representation functions as an interactive physical control. TUI attempts to embody the digital information in physical form, maximizing the directness of information by coupling manipulation to the underlying computation. Through physically manipulating the tangible representations, the digital representation is altered. In Urp, changing the position and orientation of the building models influences the shadow simulation, and the orientation of the "wind tool" adjusts the simulated wind direction.

Intangible Representation

Although the tangible representation allows the physical embodiment to be directly coupled to digital information, it has limited ability to represent change many material or physical properties. Unlike malleable pixels on the computer screen, it is very hard to change a physical object in its form, position,or properties (e.g. color, size) in real-time. In comparison with malleable "bits," "atoms" are extremely rigid, taking up mass and space.

To complement this limitation of rigid "atoms," TUI also utilizes malleable representations such as video projections and sounds to accompany the tangible representations in the same space to give dynamic expression of the underlying

digital information and computation. In the Urp, the digital shadow that accompanies the physical building models is such an example.

The success of a TUI often relies on a balance and

strong perceptual coupling between the tangible and

intangible representations. It is critical that both tangible and intangible representations be perceptually coupled to achieve a seamless interface that actively mediates interaction with the underlying digital information, and appropriately blurs the boundary between physical and digital. Coincidence of input and output spaces and real time response are important requirements to accomplish this goal.

There exist certain types of TUIs which have actuation of the tangible representation (physical objects) as the

central mean of feedback. Examples are in Touch [5], [6]. This type of force-feedback-TUI does not depend on "intangible" representation since active feedback through the tangible representation serves as the main display channel.

Key Properties of TUI

- Computational coupling of tangible representations to underlying digital information and computation.
- Embodiment of mechanisms for interactive control with tangible representations.
- Perceptual coupling of tangible representation to dynamic representations.

Genres Of Tui Applications

Tangible Telepresence

One such genre is an inter-personal communication taking advantage of haptic interactions using mediated tangible representation and control. This genre relies on mapping haptic input to haptic representations over a distance. Also called "tangible telepresence", the underlying mechanism is the synchronization of distributed objects and the gestural simulation of “presence” artifacts, such as movement or vibration, allowing remote participants to convey their haptic manipulations of distributed physical objects. The effect is to give a remote user

the sense of ghostly presence, as if an invisible person was manipulating a shared object [7] .

Tangibles with Kinetic Memory

The use of kinesthetic gestures and movement to promote learning concepts is another promising domain.

Educational toys to materialize record & play concepts have been also explored using actuation technology and

taking advantage of i/o coincidence of TUI. Gestures in physical space illuminate the symmetric mathematical

relationships in nature, and the kinetic motions can be used to teach children concepts relevant to programming and differential geometry as well as storytelling.

Tokens and Constraints

Tokens and constraints" is another TUI approach to operate abstract digital information using mechanical

constraints. Tokens are discrete, spatially reconfigurable physical objects that represent digital information or operations. Constraints are confining regions within which tokens can be placed. Constraints are mapped to digital operations or properties that are applied to tokens placed within their confines. Constraints are often embodied as physical structures that mechanically channel how tokens can be manipulated, often limiting their movement to a single physical dimension.

The Marble Answering Machine is a classic example which influenced many following research.

Interactive Surfaces – table top TUI

Interactive surfaces are another promising approach to support collaborative design and simulation which has been explored by many researchers in the past years to support a variety of spatial applications (e.g. Urp). On an augmented workbench, discrete tangible objects are manipulated and their movements are sensed bye the

workbench. The visual feedback is provided onto the surface of the workbench keeping input/output space

coincidence. This genre of TUI is also called "tabletop TUI" or "tangible workbench."

Digital Desk [8] is the pioneering work in this genre, and a variety of tabletop TUIs were developed using multiple tangible artifacts within common frames of horizontal work surface.One limitation of above systems is the computer's inability to move objects on the interactive surfaces. To address this problem, the Actuated Workbench was designed to provide a hardware and software infrastructure for a computer to smoothly move objects on a table surface in two dimensions, providing an additional feedback loop for computer output, and helping to resolve inconsistencies that otherwise arise from the computer's inability to move objects on the table.

Augmented Everyday Objects

Augmentation of familiar everyday objects is an important design approach of TUI to lower the floor and to make it easy to understand the basic concepts. Examples are the Audio Notebook [9], musicBottles [2], It is a challenge for industrial designers to improve upon a product by adding some digital augmentation to an existing digital object. This genre is open to much eager interpretation by artists and designers, to have our everyday physical artifacts evolve with technology.

i-Interactor

i-Interactor is a device which uses i-Cam, USB , i-Pen. I-Interactor is having – Cam ,which is used to capture the projected screen .Then after use of i-Pen user can directly open any file or do any operation with the computer without use of Mouse.

Acknowledgment

This paper introduced the basic concept of TUI and a variety of examples of TUI applications to address the key properties of TUI and its design challenges.

The research of TUI which gives physical forms to digital information/computation naturally crosses with the paths of industrial/product design as well as environmental/architectural design. It has also made an impact on the media arts/interactive arts community. The TUI design will contribute to promote those interdisciplinary design research initiatives in the HCI community to bring strong design culture as well as media arts perspective to the scientific/academic world.

References

[1] W.J. Book, Modelling design and control of flexible manipulator arms. A tutorial review, *Proc. 29th IEEE Conf. on Decision and Control*, San Francisco, CA, 1990, 500-506.

[2]Ullmer, B. and Ishii, H. (2000). Emerging frameworks for tangible user interfaces. IBM Systems Journal 39, 3&4, 915-931.

[3] Smith, D. (1982). Designing the Star User Interface, Byte, pp. 242-282.

[4] Johnson, J., Roberts, T. L., Verplank, W., Smith, D. C., Irby, C. H., Beard, M. and Mackey, K. (1989). The Xerox Star: a retrospective. IEEE Computer 22, 9, 11-26, 28-29

[5] Brave, S., Ishii, H. and Dahley, A. (1998). Tangible interfaces for remote collaboration and communication,

Proceedings of the 1998 ACM conference on Computer Supported Cooperative Work (CSCW 98), ACM

Press, pp. 169-178.

[6] Frei, P., Su, V., Mikhak, B. and Ishii, H. (2000a). curlybot: designing a new class of computational toys, Proceedings of the SIGCHI conference on Human factors in computing systems (CHI 2000), ACM Press, pp. 129-136.

[7] Brave, S. and Dahley, A. (1997). inTouch: A Medium for Haptic Interpersonal Communication (short paper).Conference on Human Factors in Computing Systems (CHI '97) (Atlanta, March 1997), ACM, pp

[8] Wellner, P. (1993). Interacting with Paper on the Digital Desk. Communications of the ACM 36, 7, 87-96.

[9] Stifelman, L. J. (1996). Augmenting real-world objects: a paper-based audio notebook, Conference companion on Human factors in computing systems: common ground, ACM Press, pp. 199-200.

IT 020

IMPACT OF INFORMATION TECHNOLOGY ON THE DEVELOPMENT OF RURAL ECONOMY OF INDIA

Yogesh Raaje, Abhay Chounde

Institute of Information technology, VIIT, Baramati, MS-India

yhraje@gmail.com,abhay.chounde@gmail.com

Abstract

How can information technology (IT) impact on rural economy and life of rural India to rural development? What are the challenges through which impacts can be realized, and what are the practical means for realizing potential benefits? We cannot make India economy better unless we understand the importance and strength of economy of rural sector. This paper also highlights that there is wide scope in rural market of India. This paper examines several ongoing projects that aim to provide IT 0based services to rural populations in India. These projects are distinguished by the goal of commercial sustainability, which supports scalability and, therefore, more widespread benefits. The analysis highlights the common building blocks required for successful implementation, and the relative strength and weakness of different approaches.

Keywords: ITC, eChoupal, CIC, NCAER

1. Introduction

A large number of studies have shown that today approximately 75% of Indian Population lives in Rural Areas. Today Rural Development is essential for the development of the Indian Economy. Rural economy can be developed by improving rural markets. Indian Government has realized the role of rural development and contribution of I.T. in the development of Rural Markets. A large number of projects are introduced in the rural areas with many upcoming projects in pipeline. Rural Literacy is ranked at the topmost position of priority. According to the study by NCAER

(National Council for Applied Economics Research), the number of middle and high-income households in rural India is expected to grow from 80 Million to 111 Million by the end of year 2007 as compared to Urban India that is nearby 59 Million.

The rural market of India is showing impressive growth largely due to changing lifestyle patterns, better communication network and rapidly changing demand structure of consumers of rural area. With the changing patterns of Rural Market, the role of I.T. has increased from providing only the Networks to set-up the basis of updated technological programs in the rural area. It is seen that the people living in the rural area have strong purchasing power and have more openness for new technologies as compare to the past.

In Rural India, Government has already provided kiosks, which provides the basic communication facilities like Internet connection and telecommunication services. Modern Information and Communication Technologies (ICT's) were found to have great potential to contribute. The project of Info-Kiosks are popular in th region of Andra Pradesh ,Delhi,Karnataka,Madhya Pradesh,Kerala,Gujrat,Rajasthan,Tamil Nadu and Uttar Pradesh and have shown a positive response in the development of the rural area.

1.1 Information Technology and Development

There are two types of potential economic gains from the use of IT .First, there are both static and dynamic efficiency gains. Static gains are one-time, and come from more efficient use scare resources, allowing higher consumption in the present. It is useful to distinguish two kinds of static efficiency gains. One kind pertains to increases in operating efficiency, while other comes from reduced transaction costs. In both the cases, the channel for gains is through more effective and lower cost. It is useful to distinguish two kinds of static efficiency gains .One kind pertain to increases in operating efficiency, while other comes from reduced costs. In both the cases, the channel for gains through more effective and lower cost information storage, processing and communication.

Dynamic gains come from higher growth, potentially raising the entire future stream of consumption .The second type of potential benefit comes from reduction in economic inequality, to extent that such reductions are an agreed upon social goal, and therefore social benefit. However focus on using IT for

rural development is, at least on the surface ,supportive of reduced inequality along with increased efficiency and growth .Development can also include improvements in the capabilities of population, such as education, health and nutrition, independently of any direct or indirect economic impact.

IT involves the electronic processing, storage and communication of information, where anything that can be represented in digital information is included in the term of 'information'. Thus news, entertainment, personal communications, educational material, blank and filled-out forms, announcements, Schedules and so on are all information.

Benefit of IT is increasing efficiency by economizing on resource use in the operations of firms as well as market transactions. Information that could otherwise be conveyed through the face to face contact, post, courier, print delivery, telegraph or telephone may instead be communicated in digital information form via the Internet. The ability of IT 0Based communications is to bring together buyer and sellers more effectively.

In the rural Indian context, farmers selling their crops and buying the inputs. IT makes the closeness between government agriculture exentension workers and farmers. For example farmers can get immediately any type of assistance from agriculture extension workers of agriculture department regarding when to sow the crop, when to harvest and treatment against pests and weeds and other climatic hazards through the use of IT tools like e-call centers, web portals and mobile technologies. So this process makes the effectiveness of various schemes and programmes undertaken by the government of India for the green revolution and development of rural economy of India.

1.2 Challenges and Issues for Rural IT

The power supply is the major problem in the implementing IT in rural sectors in developing country like in India. However the use of Battery backup and solar energy is the solution for the problem. Implementation of these will increase the implementation costs. Battery backups are very partial solution and solar technologies may be more promising in the near future.

2. Cost factor in Implementation of IT in Rural Sector

It is main challenge for implementing IT in rural sector in developing country like India because IT implementation includes installation of hardware components like computer machines, networking tools like routers, hubs, cables, printers and software components like operating system, and other application software. It is now possible to fully equip a single computer rural Internet kiosk is available for less than Rs.50,000,including CD Drive ,printer ,scanner, power backup and web camera. But this cost will huge when we have to install computer machines throughout the country. For this we have to raise the funds which can be done by imposing tax.

3. Issues of Awareness and Training for Using IT Tools

This is the main prerequisite in implementing IT in rural sector that we have to make awareness among rural peoples about using of IT and its benefits .Training of rural kiosk operators becomes a key aspect. Training the field personnel at various levels (village and district hub) is also critical. For this training programmes government have to take initiatives and there is also need of participation of NGO's in this direction.

4. Implemetation of IT: Various Case studies and their impacts

Government and some private sectors had introduced a number of programs through which the people of rural India can come forward and use the I.T. enabled services and work more systematically. Some of the programs run by the Government and Private sectors are:

4.1 Drishtee

Dirshtee is present in 5 States and is currently available in six districts. It is a private company, which was previously named as Cyber Edge, which has the main work of developing modules. It is present in Bihar, Haryana, Madhya Pradesh, Punjab and Rajasthan. They prepare module for poor section of the society who cannot understand the international language. The modules are designed for rural and semi-urban areas especially.Drishtee.com had its origins in Gyandoot, a government project in Dhar district of Madhya Pradesh,in central India.Dirshtee has attempted to take the model and rapidly replicate it across

the country.Currently,Dishtree has over 100 rural Internet kiosks in several states ,run by franchisees according to a revenue sharing arrangement.Drishtee is a commercial organization ,with specific social objectives of targeting benefits to the rural poor built into its vision and stratergy.Thus Drishtee's model involves not only franchising individual kiosks ,but also potentially district hubs. Partnering with local district hub 'channel partners' allows Drishtee to expand faster without creating a bulky organization, spread risks, and also insulates Drishtee from some of commercial pressure that might conflict with social objectives. It uses standard battery backup for power interruptions, and has relied mainly on dial-up internet access, though it is experimenting with Wi-Fi for district level intranets.

4.2 Rural e-seva

It was initiated by ANDRA PRADESH Government .It was initially implemented in West Godavari District to deliver e-governance facility .The centers are designed with the view to provide better governance facilities to the people of the Rural India.e-seva is gaining popularity with passing days as it helps the citizens to avail the benefit of getting the certificates at their doorsteps.

4.3 ITC

ITC stands out as a large Indian corporation serving global markets. Its kiosks are called e-choupal, and they have several differentiating features. The key distinguishing factor is that the e-choupals are totally designed to support ITC's agricultural product supply chain. In addition, the e-choupals are totally owned and set up by ITC, with the operators not having any investment or risk of their own. Management trainees are heavily impressed in the e-choupal model as part of their inculcation into ITC's working. There are four different types of e-choupals for four different products: shrimp, coffee, wheat and soyabeans.E-choupals also provide access to local market (mandi) prices and global market price information on soyabeans and derivative products, to allow farmers to compare prices. They give access operational information, developed by ITC experts, pertaining to cropping, seeds, fertilizers, and so on.E-choupals are set by ITC, with solar power backup and VSAT connectivity. The equipment cost for e-choupal is borne by ITC, with selected farmer providing the location.In addition

to the adoption advantages that come from using a farmer with high social status as the operator, the house should be specious and sturdy enough to house all the required euipement, including the VSAT and solar panel on the roof.

4.4 Community Information Centers

The program is designed especially for providing the Internet access and I.T. Enabled services to the citizens through which the interface between the Government and the Citizens can be setup. These centers connect seven northeast states namely: Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland and Tripura. The centre helps to gain the connectivity at the time of unsuitable environmental conditions .The centers are commonly termed as CIC, which are situated at the school, college or any governmental office. People can come for the Internet access, and for accessing the Internet, a nominal amount is charged from the people through which the daily expenses of centers are maintained.

5. Conclusion

Indian rural market face can be changed only with the deployment of IT. The paper has provided the description about various services offered by in the Rural areas of different states of India and simultaneously the advantages of these services. This paper has briefly surveyed several initiatives to provide IT 0based services in the rural India .The increase in the services provided to the rural people will result in the overall betterment of the society on one side by enriching the people with the updated market information and providing the latest technological development news and organizations in other side by creating more market opportunities for them. Internet based services should be provided in the rural areas, which can increase the acceptance rate of the services blended with the customer support services should be provided in the rural areas, which can increase the acceptance rate of the services by the rural people.

References

[1] Website Maintained by the Department of Information Technology of Government of India and ITC, Other websites running by the State Government of India.

[2] ACM Portal.

IT 021

Distributed Virtual Compiler Editor

Author Name: - Sameeran Tammewar

Profession: - Student (B.E. Computer)

College: - College of Engineering, Manjri, Pune-412 307

Contact Number: - 7588087618

Abstract - A web-based environment for C, Java, and UNIX shell script programming is described, together with its use in an engineering freshman C programming course. Options are availableto view files, edit source code, compile, run, run in debug mode, run with output plotted and displayed as a GIF image, display C preprocessor output, display generated assembly code, display optimized assembly code, and insert compiler error messages as comments into the source code. The environment is implemented using a combination of C code, perl, and shell scripts, and is freely available (open source). The source code of the environment itself can be used as examples in an advanced Unix/C programming or security course. The environment has been used successfully in both sophomore and senior-level C programming courses, a graduate Unix/C programming course (C and shell programming), and a senior/graduate computer communications security course (Java programming).

INTRODUCTION

The web-based programming environment described here was first presented at FIE 2004 [1] and since then has been extended in several innovative ways. The original provided a development environment for C, Java, and Unix shell script programming, and interactive I/O. Separate per-course directories and access to copy and run textbook sample programs have also been added. The implementation is open-source and available for free [2]. We will focus on the use of the environment in a freshman engineering C programming course, though it has also been used in graduate C and Korn shell programming courses, and other non-programming courses using Java for some applications.

For ECE and CSC computer programming assignments, students at Villanova have access to a variety of systems including Unix and Microsoft workstations in departmental and college laboratories, as well as their own personal computers. But the programming environments available on these systems vary widely – some may not have Java installed, or may have an unsuitable old version; some only have professional C++ development tools with complex interfaces that overwhelm beginners and do not enforce strict ANSI C compliance; and the Microsoft systems generally do not have the Unix shells and other tools (e.g. Cygwin [1]) installed. To alleviate this problem, the View/Edit/Compile/Run (VECR) environment [2] was created to provide an easy-touse interface to a set of standard programming tools, with consistent options. For example, for C programming, the GNU GCC compiler [3] is used with options *–ansi –pedantic –Wall* to enforce strict ANSI C compliance and enable all warnings. This helps students to learn standard C and avoid use of system-specific functions or inappropriate constructs from C++.

Students are encouraged to use whatever programming environment is most convenient for them for initial development of programs, then upload to VECR for testing and turning in. However most students have reported that they do all development directly in VECR, after having used it and found it to be both convenient and consistent, providing the same interface for C, Java, and Bourne shell programming.

During the process of software development frequently more than one compiler package is required. Some products are known to be very useful for locating errors or debugging, while others perform extremely well when a program or library is in the final stage of development and should be optimized as much as possible. Also when facing obscure error messages, which may result in a time-consuming search for the error, a different error message from the second compiler frequently cuts that time dramatically.

Therefore students should be to some extend exposed to different compilers at some point in their software courses curriculum. Although all necessary software is installed in the computer laboratories, most students prefer to work on their computers at home or dormitory and connect to the university network. That situation creates unnecessary burden either for the network administrators who have to install additional software on many machines of non-standard configuration, or on students who must purchase and install on their own several software packages along their full course of study.

Intranet Compilers Architecture:-

During software development it is important to justify which part of the software should run on the client machine and which part should run on the server.

Client side programs - applets are transferred through network when requested and execution is performed entirely on the client machine that made the request. This allows for sharing the computational cost between the server and client. It approach can be used when programs to be transferred to users are moderate in size or are be cached on client machine, or the data to be transferred between server and client, in case the application is run on the server, are very large in volume. In case of platform independent solutions, such as Java, causing lesser computational performance may be prohibitive.

With CGI much less information has to be passed to the server. The server executes instructions based on the given information and sends the results back to the local machine that made the request5. This is used in the opposite cases, when the software package is large or should not be released to user, or when amount of data to be transferred is small. However, large number of clients that access the server simultaneously would make CGI-based approach undesired.

These software design problems were considered and solved in the ICP. The user interface is programmed in HTML enhanced with JavaScript. The purpose of the project was allowing students to get familiar with different compilers and compiler optimization techniques rather than make another huge GUI application to wrap compilers. Therefore, it is assumed that the user will use his or her favorite text editor to create and correct program files. This assumption allowed to create the very simplified front-end that loads quickly and is really platform independent.

The server side part of the application is implemented using CGI scripts written in PERL that handle communication between a user and different compilers. That script does the file managing, runs compilers and processes the compilation results6. The result is both the source code listing and a binary code to download or a list of errors sent back to the user.

To use ICP, paste the program code from your compiler text editor, or from any text editor, to the web page form. Then submit the form. The compilation will be performed by PERL script on the server in batch mode. Although the front end is designed to be as simple as possible with only a few commonly used options, it is sufficiently functional and can be used quickly. The PERL script located on the server has to deal with the translation of these common options to the actual options of compilers from different vendors. It also handles the compilation errors and processes the report.

Intranet Compilers User Interface

HTML page allows for selecting a vendor and a language, and for setting a few basic compilation options. User uses copy and paste commands to enter the

source code into the compiler front end. There are three menus located under the source code area as shown.

The middle menu is used to select the programming language while right menu is used to select the compiler vendor. Currently the Intranet Compilers package supports C, C++, Pascal, Fortran, and Java languages. It utilizes DJ GNU (ver. 2.91), Borland (ver. 4.5), and Microsoft (ver. 5) compilers for C and C++, GNU compiler for Fortran and Pascal, and Sun's JDK for Java. At this moment partial compilation into assembly language is supported only for GNU and Borland compilers.

One of preset compiling configuration can be selected from the left menu. User can decide whether aggressive binary code optimization or strict error checking and ASNI style violation checking are necessary. The compiler vendor and version can be selected from the right menu. In case of selecting one of commercial compilers while working at off-campus location user is requested to input a password to verify his or her legibility to use licensed products.

One of the major advantages of consolidating more than one compiler is the ability to cross-reference error message among different vendor products used from the same interface.

The process of compilation is performed in batch mode. After setting the desired options and pasting the source code into the appropriate text box, the task can be started by pressing the

Compile button. As a result another web page with HTML wrapped information is sent back to the user. The result page is displayed in another browser window so that the user can correct the source code in the original window and resubmit it if necessary.

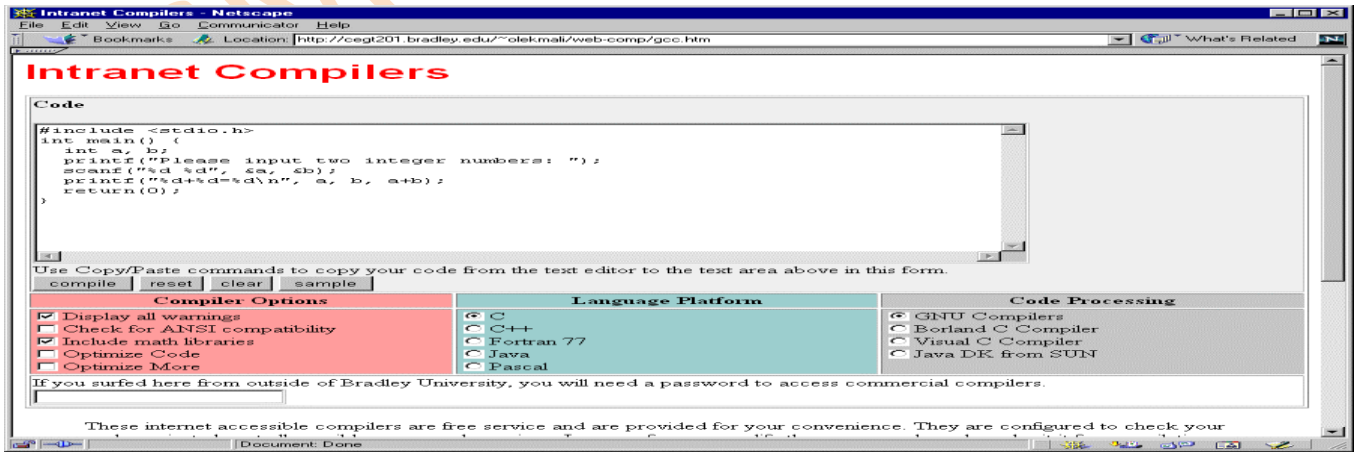


Fig. 3. The common web-based front-end to C++ compilers

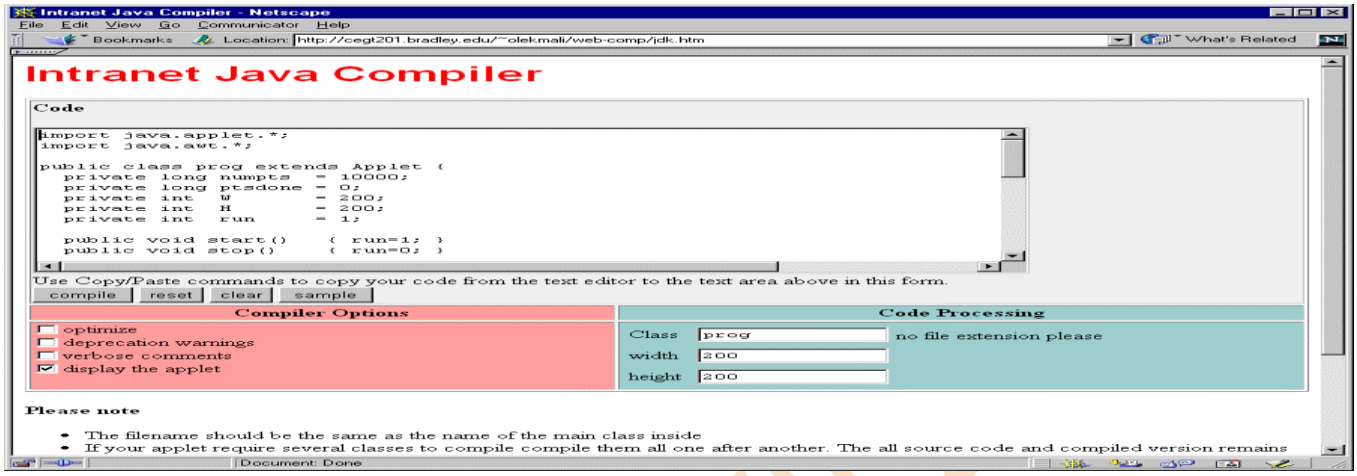


Fig. 4. Specialized front ends for Java

Fig. 5. result of Fig.4.

After successful compilation the binary code is available for download. If the user's operating system agrees with the destination platform set for the compiler the binary code may be executed with full user interaction.

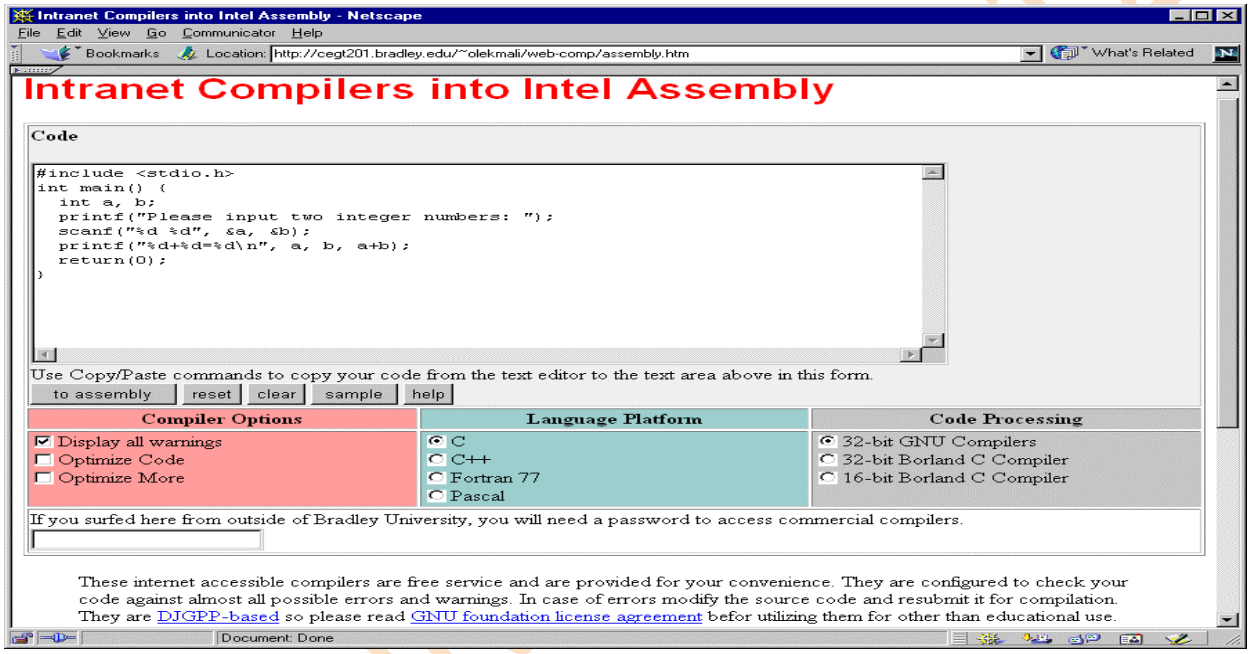


Fig. 6. Specialized front ends for assembly language generator

Fig. 7. result of Fig.6.

Because of additional requirements for displaying Java applets, an additional dedicated front end is created for that language. It is shown in Fig. 4. If the compilation is successful, the user can inspect the applet immediately. Another specialized page was created for the partial compilation into assembly language because users are interested in different set of compiler options that would be useful for generated assembly language code inspection. 16- and 32-bit Intel assembly languages are available. However, a cross-compiler may be used to produce assembly language code for different machines. The assembly front-end and the result page are shown in Fig. 6 & 7.

Java Online Compiler

The JavaOnline compiler is another component of JavaOnline developed using Java SE 6 (Sun Microsystems Inc, 2009c). This component runs independent from the Moodle server. It is a JAR file which is deployed on the web server. On the start up of the web server, this Jar file gets initialized and checks the path detail where students programs get saved. Java SE 6 (Sun Microsystems Inc,

2009c) is the current major release of the Java SE platform. In this release they have introduced new package javax.tool that offers interfaces for retrieving and compiling files passed as strings. Thus clients can locate and run Java compilers from within a program. The package also pro-vides other interfaces that generate diagnostics and override file access in order to support the compilation process (described below). Figure 8 shows javax.tool interface.

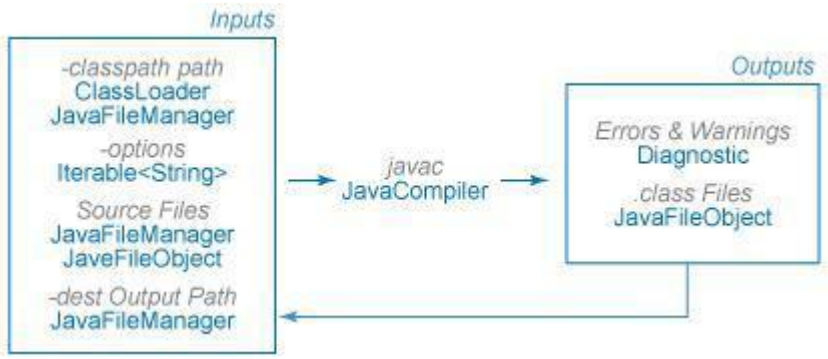


Figure 8. javax.tool Interface

Java Compiler Interface

The new JavaCompiler interface of the tool package invokes Java programming language compilers from programs. It is JSR 199 specification released in final form (Peter von der Ahé, 3 April, 2007). The interface can create Java classes by compiling valid Java source, bypassing the previous need to validate byte code, or understand new object models of classes, methods, statements and expressions (David J.Biesack, Dec, 2007). It simplifies and standardizes the compilation process by providing one supported mechanism for code generation and loading that is not limited to passing a single file of source code. The JavaCompiler relies on two services: DiagnosticListener and JavaFileManager to generate diagnostics during compilation. The following example shows creation of JavaCompiler object.

```
/*Initialize Java Compiler Class */

JavaCompiler compiler = ToolProvider.getSystemJavaCompiler();
```

Diagnostic Interface

The Diagnostic Interface usually reports a problem at a specific position in a specific file. It supplies error details like the line number, column number, start position, end position, error message, source of error and kind of error. It also provides details about warning and information messages generated during compilation. The diagnostic listener which is part of diagnostic interface that listen to error messages and it will be written to default output by means of System.err. There is another class called diagnostic collector that has java bean properties which get initialized when diagnostic listener finds errors. The following example shows creation of the DiagnosticCollector.

```
/* Create Diagnostic Collector to store compilation Errors */

DiagnosticCollector<JavaFileObject> diagnostics = new DiagnosticCollector
 <JavaFileObject>();
```

File Object Interface

The File Object Interface is an abstraction tool. In this context file means abstraction of regular files and other sources of data. For example, a file object can be used to represent regular files, memory cache or data in database. All methods in this interface might also throw a security Exception if security rules are violated. The StandardJavaFileManager interface serve two purposes: it reduces the overhead of scanning files and reading Java archive (JAR) files. The following example (Sun Microsystems Inc, 2008) shows a coding pattern.

```
/* Initialize File Manger to supply path of Java file*/

StandardJavaFileManager fileManager = compiler.getStandardFileManager(
 file,diagnostics, null, null); compiler.getTask(null, fileManager, diagnostics, null,
 null, null).call();

for (Diagnostic diagnostic : diagnostics.getDiagnostics())

 System.out.format("Error on line %d in %d%n", diagnostic.getLineNumber()
 diagnostic.getSource().toUri()); fileManager.close();
```

File Execution

On successful program compilation, Java compiler generates byte code in the form of class file. This file contains executable code in the form of instruction

that Java virtual machine executes. The JavaOnline plug-in then uses Runtime.exec() method to execute. On some of the platforms, Runtime.exec() throws stackoverflow exception because of limited output buffer size. To avoid this situation, JavaOnline itself provide input and output buffers. On the execution, output gets stored in these buffers and passed back to the user interface. On the execution error, the JavaOnline fetches errors from System.err to buffer and send back to User Interface. The following example shows creation of new process using Runtime.exec() method.

```
Runtime rt = Runtime.getRuntime();
Process proc = rt.exec("java classfilename");
 InputStream stderr = proc.getErrorStream();
 InputStreamReader isr = new InputStreamReader(stderr);
BufferedReader br = new BufferedReader(isr);
String line = null;
while ( (line = br.readLine()) != null) System.out.println(line);
int exitVal = proc.waitFor();
 System.out.println("Process exitValue: " + exitVal);
```

Conclusion

The paper shows how to use the new opportunity created by Internet technologies for the efficient and platform independent usage of software engineering tools. The presented ICP are just an example, but it shows a way in which the technology can be implemented. The possibility to compare error messages from different compilers is very useful especially for students who do not have much experience in coding and debugging yet. The ability to use different compilers allows a programmer to pick up the fastest or the most convenient tool to compile the code and remove the errors. It also offers an easy way to analyze the results of compilation by inspection of generated assembly code.

The Java Online question type plug-in for Moodle LMS presented is presented in this document. This provides a key solution for online compilation of Java source code. This plug-in enables students to compile their programs without having to configuring their machine for Java program compilation. This also allows students to do the Java programming online, anywhere, anytime. The

provided detailed diagnostics and log helps students to find compilation and execution errors much easily.

Bibliography

1. Sweet, W. and Geppert, L., "http:// It has changed everything, especially our engineering thinking," IEEE

Spectrum, January 1997, pp. 23-37.

2. Camposano, R.; Deering, S.; DeMicheli, G.; Markov, L.; Mastellone, M.; Newton, A.R.; Rabaey, J.; Rowson,

3. J.; "What's ahead for Design on the Web", IEEE Spectrum, September 1998, pp. 53-63.

4. Blackboard Inc, 2009, Blackboard home, viewed March 21, 2009, http://www.blackboard.com/.

5. Douglas Kramer, 1996, The Java Platform: A white paper, viewed March 23, 2009, http://java.sun.com/docs/white/platform/javaplatformTOC.doc.html

6. Hank Shiffman, Making Sense of Java, http://www.disordered.org/Java-QA.html

7. Hank Shiffman, Boosting Java Performance: Native Code and JIT Compilers,

http://www.disordered.org/Java-JIT.html

8. Gundavaram, S.,. CGI Programming on the World Wide Web. OReilly & Associates, Inc., 1996.

9. Wall,L., Christiansen, T., Schwartz, R.L. Programming Perl, OReilly & Associates, Inc., 1996

IT 022

CHALLENGES IN INTERNET TRAFFIC MANAGEMENT

Mrs. Minaxi Doorwar , Ms. Archana Walunj

minaxirawat@rediffmail.com archana.walunj22@gmail.com

G.H.Raisoni.College of Engineering and Management Wagholi,Pune

ABSRTACT

One of the major research challenges in the design of the future Internet is to include management into the network. To serve this purpose, many researchers believe that the task decomposition capability is one of the important requirements for the management of the future Internet.In the Internet today, traffic management spans congestion Control (at end hosts) , routing protocols(on routers),and traffic engineering. Historically, this division of functionality evolved organically. a top-down redesign of traffic management using recent innovations in optimization theory. an objective function that captures the goals of end users and network operators . In todays Internet we observe the growing trend for services to be both provided and consumed by loosely coupled value networks of consumers, providers and combined consumer/providers. As networks grow in size and complexity network management has become an Increasingly challenging task. Using all known optimization decomposition techniques, generate four distributed algorithms that divide traffic over multiple paths based on feedback from the network links. Many protocols have tunable parameters, and optimization is the process of setting these parameters to optimize an objective. In recent years, optimization techniques have been widely applied to network management problems.

1.INTRODUCTION

Traffic management is the adaptation of source rates And routing to efficiently use network resources .Traffic Management has three players :users ,routers, and operators .In today's Internet, users run congestion control To adapt their sending rates at the edge of the network. Inside a single Autonomous System(AS),routers run shortest-path routing based on link weights. Operators

tune link weights to minimize a cost function [1]. Network management is the continuous process of monitoring an network to detect and diagnose problems, and of configuring protocols and mechanisms to fix Problems and optimize performance. Traditionally, network management has been Largely impenetrable to the research community since many of the problems appear Both complex and identified. In the past few years, the research community has Made tremendous progress casting many important network management problems As optimization problems. Network optimization involves satisfying network management objectives by setting the tunable parameters that control network behavior. Solving an optimization n problem involves optimizing an objective function subject To a set of constraints. Unfortunately, while convex optimization problems are easier To solve, many problems that arise in data networks are non convex. Consequently, they are computationally intractable, with many local optimal that are suboptimal. The design of optimizable networks network architecture sand protocols that lead To easy-to-solve optimization n problems and consequently, optimal solutions.The changes to protocols and architectures can range from minor extensions to clean- slate designs. In general, the more freedom we have to make changes, the easier it would be to create an optimizable network. On the other hand, the resulting improvements in network management must be balanced against other considerations such as scalability and extensibility, and must be made judiciously. Tomake design decisions, it is essential to quantify the trade-off between making network-Management problems easier by changing the problem statement and the extra over- Head the resulting protocol imposes on the network.[2] Rethink Internet traffic management Using optimization theory as a foundation. Optimization decomposition is the process of decomposing a single optimization problem in to many sub-problems, each Of which is solved locally. The barriers are two- fold. First, any mathematical modeling make simplifying assumptions. Second, while multiple decomposition methods exist, it is unclear how to compare them.

1.1 Protocol Design using Decompositions:

Demonstrate how to create a practical network protocol by deriving multiple distributed algorithms, comparing the practical properties, and synthesizing their best features in to a practical protocol.

1.2 Redesigned Traffic Management:

introduce TRUMP, a Traffic-management Using Multiple Protocol to replace congestion control and Traffic engineering. TRUMP is easy to manage and robust

to small-time scale traffic shifts. TRUMP converges faster than the four algorithms and has the fewest tunable parameters. As with any mathematical modeling, the TRUMP algorithm leaves many protocol details un specie. Use out intuition to address the details. The TRUMP protocol is evaluated Using packet-level simulations with realistic topologies .[1]

2. TODAY's TRAFFIC MANAGEMENT

In this section, introduce how optimization is used in the context of traffic management inside a single Autonomous System(AS).Traffic management has three Players :users, routers, and operators. In today's Internet, users run TCP congestion control to adapt their sending rates at the edge of the network based on packet loss. Congestion control has been reverse engineered to be implicitly solving an optimization problem, Inside the network, operators tune parameters in the existing routing protocols to achieve some network-wide objective in a process called traffic engineering,seeFigure1.

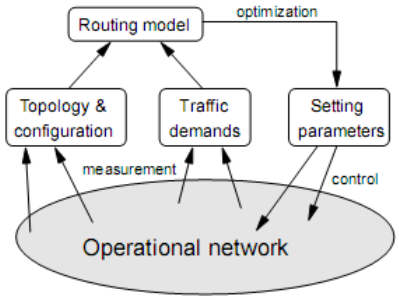


Fig.1 Components of the route optimization frame work.

2.1TrafficEngineering

Symbol	Meaning
(i, j)	Pair of routers.
$x^{(i,j)}$	Traffic demand between i and j.
l	A single link.
w_l	Link weight l.
c_l	Capacity of link l.
y_l	Traffic load on link l.
$f(y_l/c_l)$	Penalty function as a function of link utilization.
$r_l^{(i,j)}$	Portion of the traffic from router i to router j that traverses the link l.

Table1 Summary of notation for Section2.1.

Inside a single AS, each router is configured with an integer weight on each of its outgoing links, as shown in Figure2.The routers flood the link weights throughout The network and compute shortest paths as the sum of the weights.

For example, i Directs traffic to k though the links with weights(2,1,5) Each router uses this information to construct a table that drives the forwarding of each IP packet to the next Hop in its path to the destination. These protocols view the network inside an AS is a graph where each router is a node $n \in N$ and each directed edge is a link $l \in L$ between two routers. Each unidirectional link has a fixed capacity c_l ,as well as a Configurable weight w_l .The outcome of the shortest-path computation can be represented as $r_l^{(i,j)}$:the proportion of the traffic from router i to router j that traverse The link L. Operators set the link weights in intra domain routing protocols in a process called traffic engineering. [2]

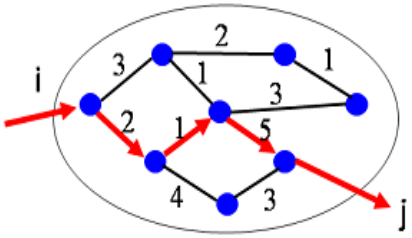


Figure.2 Network topology with link weights for shortest path routing.

The selection of the link weights w should depend on the offered l trraffic, as captured by a demand matrix whose entries x represents the rate of traffic entering at router i that is destined to router j. The traffic matrix can be computed based on traffic measurements or may represent explicit subscriptions or reservations from users. Given the traffic demand $\Sigma_l f(y_l/c_l)$ and link weights w ,the volume l of traffic on each link l is $y_l = \Sigma_{i,j} x^{(i,j)} r_l^{(i,j)}$,the proportion of traffic that traverses link l summed over all ingress-egress pairs. An objective function can quantify the "goodness" of a particular setting of the link weights. For traffic engineering, the optimization considers a network-wide objective of minimizing $\Sigma_l f(y_l/c_l)$.The traffic Engineering penalty function f is a convex, non-decreasing, and twice-differentiable Function that gives an increasingly heavy penalty as link load increases, such as an Exponential function. The problem traffic engineering solves is to set link weights $\Sigma_l f(y_l/c_l)$, assuming the weights are used for shortest-path routing.

3. CHOOSING AN OBJECTIVE FUNCTION

Every optimization problem consists of an objective function, constraint set and variables. For traffic management, by having both routing and source rate as optimization variables, we have the most flexibility in re- source allocation. In our problem, the constraint is that link load does not exceed capacity. The objective function remains to be designed. To address practical challenges select an objective which balances maximizing user utility with minimizing operator's cost function.

3.1 Maximizing Aggregate Utility: DUMP

One natural objective for the traffic management sys tem is to maximize aggregate user utility , where

$U_i(x_i)$ is a measure of "happiness" of source-destination as function of total transmission rate x_i. U is concave non negative, increasing and twice deferential function i.g $\log(x_i)$ that that can also represent the elasticity of the traffic.

$$\begin{aligned} &\text{maximize} && \textstyle\sum_i U_i(x_i) \\ &\text{subject to} && \mathbf{Rx} \preceq \mathbf{c},\ \mathbf{x} \succeq \mathbf{0} \end{aligned} \qquad (1)$$

Where both R and x are variables.

A distributed solution to(1) can be derived through dual decomposition if (1) is a convex optimization problem. To capture multipath routing, introduce z to represent the sending j rate of source i on its jth path . Also represent available paths by a matrix H where

$$H_{lj}^i = \begin{cases} 1, & \text{if path } j \text{ of source } i \text{ uses link } l \\ 0, & \text{otherwise.} \end{cases}$$

H does not necessarily present all possible paths in the Physical topology, but a sub set of paths chosen by op- Erators or the routing protocol .Then we can rewrite (1) as:

$$\begin{aligned} &\text{maximize} && \textstyle\sum_i U_i(\sum_j z_j^i) \\ &\text{subject to} && \textstyle\sum_i \sum_j H_{lj}^i z_j^i \le c_l,\ \ \forall l. \end{aligned} \qquad (2)$$

In this form,(2) is a convex optimization problem. A Distributed solution to (2)can be derived using dual decomposition ,where a dual variable is introduced to relax the capacity constraint. The resulting Dual-based Utility Maximizing

Protocol(DUMP) involving sources And links is summarized in Figure1.Similar to the Reverse engineering of the congestion-control protocol in, s can be interpreted as link prices.

Feedback price update at link l:

$$s_l(t+1) = \left[s_l(t) - \beta_s(t) \left(c_l(t) - \sum_i \sum_j H^i_{lj} z^i_j(t) \right) \right]^+ ,$$

where β_s is the feedback price step size.

Path rate update at source i, path j:

$$z^i_j(t+1) = \text{maximize}_{z^i_j} \left(U_i \left(\sum_j z^i_j \right) - z^i_j \sum_l s_l(t) H^i_{lj} \right)$$

Figure 3 the DUMP algorithm

DUMP is similar to the TCP dual algorithm when the step size is too large, DUMP will constantly over shoot or undershoot, never reaching the ideal utility. On the other hand, when the step size is to o small, DUMP converges very slowly. Even at the optimal step size, DUMP only converges after about 100 iterations. This highlights that choosing an appropriate step size for DUMP is challenging.

3.2 New Objective for Traffic Management

Let us reflect for a moment on why DUMP has poor convergence behavior. If we look at the form for feed- back price, we see it is only nonzero when links are overloaded, therefore, the feedback from the links is not fine-grained. In fact, the feedback price in DUMP has the same formulation as the congestion price [5].To avoid the poor convergence properties of DUMP we look for an alternative problem formulation which also takes into account the operator's objective. To-day, traffic engineering solves the following optimization problem with only R as a variable (and x constant):

$$\text{minimize} \quad \sum_l f\left(\sum_i R_{li} x_i / c_l\right). \qquad (3)$$

f is a convex, non-decreasing, and twice-deferential function that gives increasingly heavier penalty as link.[3] If solve(3) with A better traffic management objective could be to combine performance metrics (users' objective) with net-work robustness (operator's objective), leading to the following

formulation as a joint optimization over(x,R):

$$\begin{array}{ll}\text{maximize} & \sum_i U_i(x_i) - w\sum_l f(\sum_i R_{li}x_i/c_l) \\ \text{subject to} & \mathbf{Rx} \preceq \mathbf{c},\ \mathbf{x} \succeq \mathbf{0}.\end{array} \quad (4)$$

This objective favors a solution that strikes a trade-o between high aggregate utility and a low overall network congestion, to satisfy the need for performance and robustness. Similar problem formulations were proposed Before generating distributed solutions we first transform (4) to a convex optimization problem:

$$\begin{array}{ll}\text{maximize} & \sum_i U_i(\sum_j z_j^i) - w\sum_l f(y_l/c_l) \\ \text{subject to} & \mathbf{y} \preceq \mathbf{c}, \\ & y_l = \sum_i \sum_j H_{lj}^i z_j^i,\ \ \forall l.\end{array} \quad (5)$$

Note that to decouple the objective which contains U (a per-source function)and f (a per-link function),we Introduce an extra variably to provide feedback before l link load exceeds actual capacity.[1]

4.TRUMP

While the algorithms introduced in Section 3 con verge faster than DUMP, we seek an algorithm with even better convergence properties.

Feedback price update:

$$s_l(t+1) = p_l(t+1) + q_l(t+1),$$

Loss price update:

$$p_l(t+1) = [p_l(t) - \beta_p(c_l - \sum_i \sum_j H_{lj}^i z_j^i(t))]^+,$$

Delay price update:

$$q_l(t+1) = wf'\left(\sum_i \sum_j H_{lj}^i z_j^i(t)/c_l\right),$$

Path-rate update:

$$z_j^i(t+1) = \text{maximize}_{z_j} U_i\left(\sum_j z_j^i\right) - \sum_l s_l(t) \sum_j H_{lj}^i z_j^i$$

Figure 4. TRUMP algorithm

Traffic-management Using Multipath Protocol (TRUMP) with only one easy to tune parameter TRUMP is a heuristic .To prove the convergence of TRUMP when the network is lightly loaded o TRUMP is simpler than any of the algoriyhm with only

one tunable parameter that only needs to be tuned for small w.

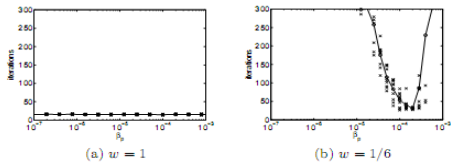


Figure 5: Plots of TRUMP algorithm showing dependence of convergence time on step- size. 'x' represent the actual data points and 'o' represent the average value.[4]

CONCLUSION

A traffic-management protocol which is distributed, adaptive, robust, flexible and easy to manage. TRUMP is easy to manage, with just one optional tunable parameter. using optimization decompositions as a foundation, simulations as a building block, and human intuition as a guide can be a principled approach to protocol design. With the increase in urbanization and traffic congestion, greater demand is placed on operating traffic systems with maximum efficiency[6]. The intelligent traffic management system proposed in this work is a distributed automation systems based on Internet and Web technologies.[7] The system uses the Ethernet as a communication backbone between individual nodes located at different traffic intersections and a central traffic management unit.

REFERENCES:

1) JiayueHe, MartinSuchara, Maayan Bresler, "Rethinking Internet Traffic Management: From Multiple Decomposition to practical protocol"IEEE Jan 2007
2) Jiayue He ,Jennifer Rexford and Mung Chiang "Design forOptimizability: Traffic Management of a Future Internet" IEEE Jan 2009.
3) Raida Al-Alawi, Member IAENG "Web-Based Intelligent Traffic Management System" IEEE 2008.
4) Jennings, B. Brennan, R. Donnelly, W. Foley, S.N. Lewis, D. O'Sullivan "Challenges for federated, autonomic network management in the Future Internet " IEEE August 2009 .
5) Augustin Soule, Haakon Larsen, Fernando Silveira, and Christophe Diot " Detectability of Traffic Anomalies in Two Adjacent" IEEE 2009.
6) R. Srikant, The Mathematics of Internet Congestion Control. Birkhauser, IEEE 2004.
7) J. He, M. Bresler, M. Chiang, and J. Rexford, "Towards multi-layer traffic engineering: Optimization of congestion control and routing," IEEE June 2007.

IT 023

SYSTEM DEDICATED TO PROCESS SINGING VOICE FOR MUSIC RETRIEVAL IN INDIAN CLASSICAL TERMINOLOGY

Jui Jamsandekar, Mayuri Karle PVG's COET,PUNE-9

Abstract— In this paper ,Frequency-Pitch processing system is presented, oriented to produce Notations for a vocal tune sung by the vocalist . The system is built using MATLAB software. The paper represents the concept of pitch, timber and some terminology related to Indian classical music and terms related to human singing. This paper makes a good compliment of technology and musical science by analyzing human voice and frequency pattern. For analysing human voice Sampling, Pitch determination techniques are presented it the paper.

***Keywords*—Frequency domain, Hindustani music, Pitch Contour, Quantization.**

INTRODUCTION

In the last few years, a number of music retrieval systems have been presented in consequence of the growing amount of digital music available on the Internet. Methods for searching or classifying music have been investigated in order to improve both the tiling process on the database side and the quality of information retrieval on the client side. Trying to exploit the multimodal nature of music, methods for indexing and querying music data by content and plotting the musical notes of a singing voice in Indian classical manner are the subject of an increasing number of studies. In a wide sense, music retrieval systems should add content based interfaces next to text-based one . Music should be not only searched by textual attribute (e.g. song title) but we

should also be able to represent the notation of a song sung by a singer in the Indian classical as well as western language. In the latter case, the preferred strategies have been processing the given sound input , where the input is entered as electronic music score or by singing . In particular, System interfaces should approach even non-professional users to use musical notation retrieval system with ease. This kind of system usually adopts naive signal processing algorithms to extract pitch information from an audio input given by the user and some approximate methods for plotting thw musical similarities in a database of symbolic music files (e.g. MIDI files).

Preliminary study will be focused on the first aspect for which dedicated methods are needed because of the peculiar facets of the human singing voice . Thus, addressing the problem of converting an acoustic input i.e. Human voice into note-like attributes is becoming a central issue for correct singing as well as for the traditional and difficult task of automatic music transcription. The process could involve several stages: note segmentation, pitch tracking and note labelling. The first two stages fall in the typical domain of signal processing and techniques originally developed for speech analysis can be successfully employed. The third stage deals with the music content of the audio signal. For each note event, an estimation of pitch measured in Hertz must be converted into note labels; this process should be carried out with respect to a musical scale, for example the equal tempered scale that breaks the octave into twelve units known as semitones. This procedure is critical because intonation is never absolute and singers always introduce errors during a real performance. We stress the importance of pre-processing the audio signal in order to improve pitch detection. Preliminary bservations on processing voice in adverse conditions are also taken under consideration. The novelty of a schema for automatically adjusting user's intonation has to be developed in the system. We will give a brief overview of our system in preliminary Project report, referring the interested reader to the cited work for details.

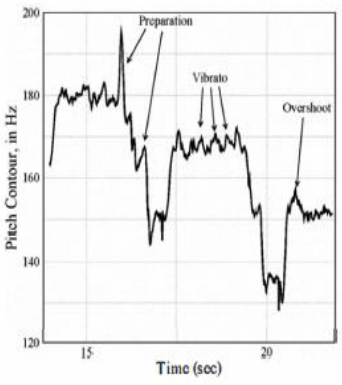


Fig. 1: A typical pitch contour of actual singing voice

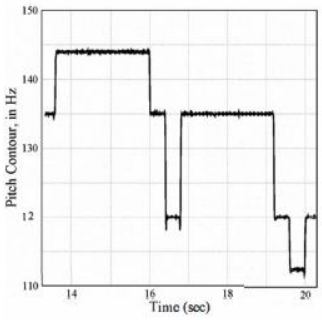


Fig. 2: The synthesized discrete pitch contour for the song used in Fig. 2

II ACOUSTIC FEATURES TO BE CONTROLLED

It has been found that the conversion from a neutral to singing voice can be done by controlling two factors, namely, duration and pitch contour of the speech [].

Dumation

In singing, the vowels are often sustained for much longer duration as compared to that in speech. More essentially, singing voice duration is constrained by the rules of rhythm, as explained below.

- The time intervals for which the notes are to be played is defined by rhythm.
- The tala is the rhythmic structure, also called time cycle. This is supposed to remain fixed within each composition, however, it can be repeated in cycles and each cycle can be divided into parts, either equal or unequal. In western music, each segment (vibhag) has equal number of beats (matras), whereas in Indian music, the number of matras per vibhag may not be equal.

Pitch (FO) contour

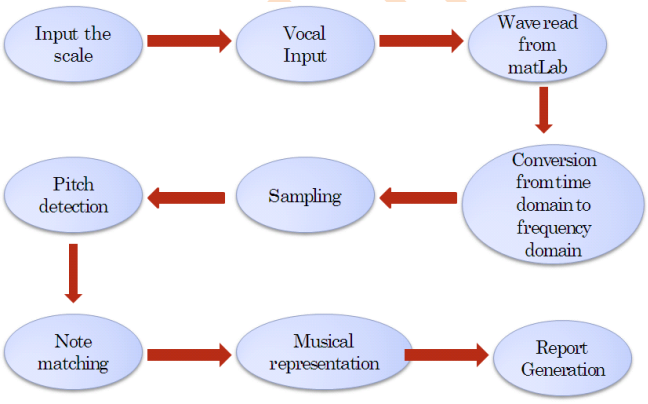
The singing voice pitch contour should be controlled in terms of global and local FO variations. Global FO variations correspond to the sequence of musical notes and local FO variations include FO fluctuations. The discrete pitch contour (Fig2) represents the global FO variations and it has to be superimposed with local FO variations to get the final singing pitch contour. There are five kinds of local variations, namely: overshoot, undershoot, vibrato, fine-fluctuation, and preparation, as having significant effect on perceived naturalness of the singing

voices in western music, by various psychoacoustic experiments. Figures 1 and 2 compare the pitch contours of real and discrete synthesized singing voices to identify some of the local variations. Most importantly, to implement the concept of musical curves, as in Hindustani music, we need to design an algorithm to introduce various ornaments in music, which are actually very essential part of Indian classical music []. The various ornaments are:

1. Meend is the connection of two notes without a break. While progressing from the starting to the ending note, it gently touches every microtone in between them, although not distinctly heard. It is different from legato in western music .
2. Gamak is when a note is shaken very heavily so that it goes higher or lower than the desired note.
3. A ndolan is a simple shaking of a note. It is like vibrato with slower speed.
4. Kana is when notes are jerked very quickly, giving hints of notes yet to come. It is delicately touching a note before finally proceeding to it or another note. Apart from duration and pitch contour, acoustic features like spectrum, amplitude etc. also have significant effect on quality of singing voice, but this paper does not deal with them.

III PROPOSED METHOD

The block diagram of the proposed conversion system, comprising of models controlling the various acoustic features of singing voice, is shown in figure . The input to the system-Human Voice and the output will be Musical Notation and vocal Analysis Report.



IV. Unit

To convert pitches from Hertz to a music representation,we use the definition of MIDI scale (see Equation below**; fl** is **8.1758** ***Hz).***

Note(md)=(1/Log2^(1/12))(Log f1/f0)

ALGORITHMIC STRATEGIES STUDIED

Pitch tracking algorithms and methods are the important part and the base of mathematical model of our system .Analysing the system mathematically for two methods of pitch tracking techniques:

1.ZCR
2.Auto correlation

Zero-crossing Rate (ZCR)

Classical methodology: The zero-crossing rate (ZCR) is the rate of sign-changes along a signal, i.e. the number of times that the signal crosses the 0 (zero) level reference. The thought is that the ZCR should be directly related to the number of times the waveform repeats itself per unit time.

Let *s(n)* be a discrete signal, of length *N*. The ZCR of this signal is found as follows:

$$ZCR = 1/N \sum_{n}^{N-1} \Xi \{S(n).S(n-1) < 0\} \quad \ldots\ldots(1)$$

The operator $\Xi \{A\}$ is 1 if the argument *A* is true and 0 otherwise. The frequency of the signal (*FSIGNAL*) is given by the following expression.

F SIGNAL = zcr. Fs / (2) .....(2)

In the equation (2), *zcr* is the ZCR calculated in (1) and *fs* is the sampling frequency of the signal. This technique is very simple and inexpensive but is not very accurate. In fact, when dealing with highly noisy signals or harmonic signals where the partials are stronger than the fundamental or if the signal has oscillations around zero axis (musical notes show both characteristics), the method has poor results.

Proposed methodology for musical notes:

Figures illustrates the waveform of the musical notes C(2) (upper) and G(1) (lower) - see table I. It is easy to see that the direct application of equation (1) will generate false results since periodic oscillations are present. However, it is possible to apply the equation (1) from a determined threshold. In order to find the threshold proposed in this work, the signal is rectified and its mean is calculated. Moreover, a security margin of 20% is used. This value is based on practical tests performed with different signals. Thus, if *s(n)* is the analyzed signal, the threshold *L* proposed is given by:

$$L = 1.2 * 1/N \sum |s(n)| \quad \text{.....(3)}$$

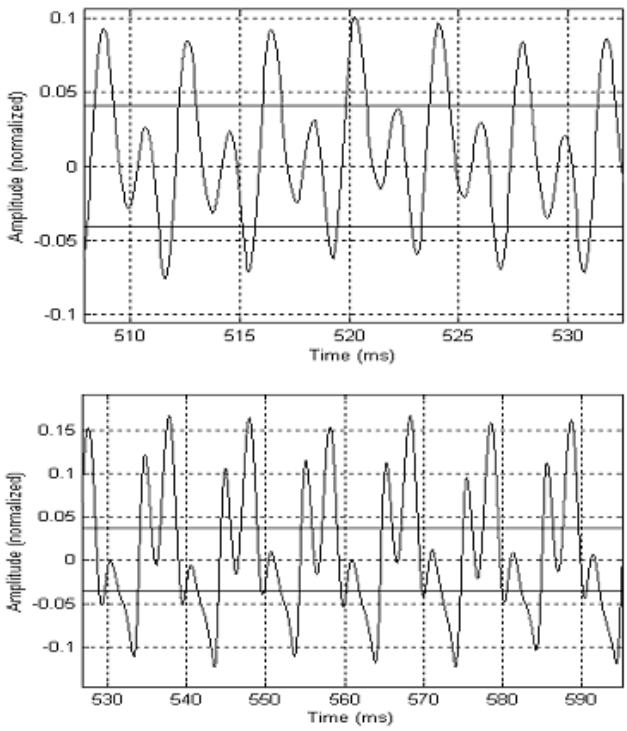


Figure 1. Waveform of the note C(2) (upper) and G(1) (lower) illustrating the thresholds used in the assessments.

Considering the proposed threshold, it is important to assess the result to positive and negative displaced versions of the original signal. The final result will be the mean of both assessments separately. To measure the ZCR considering a positive displacement of the original signal, a new signal *sp(n)* is used, which is given by:

$$Sp(n) = S(n) - L \quad 0 \le n \le N-1 \qquad \ldots\ldots\ldots (4)$$

Another improvement used consists in counting just negative-to-positive transitions. Thus, the expression must be rewritten as follows:

$$ZCRp = 1/N \sum_{n=0}^{N-1} \Xi \{Sp(n-1) < 0 < Sp(n)\} \ldots (5)$$

This new characteristic of computing just negativeto- positive transitions demands that equation (2) be rewritten as presented in (6), because each transition corresponds to one period of the signal. (In expression (2), each transition corresponded to a half period).

$$F\ SIGNAL(p) = ZCRp * Fs \qquad \ldots\ldots\ldots (6)$$

To measure the ZCR with a negative displacement, another new signal *sn(n)* is used, which is given by:

$$Sn(L) = S(n) + L \qquad \ldots\ldots\ldots (7)$$

As proposed in equation (5), the ZCR with negative displacement is given by:

$$ZCRn = 1/N \sum \Xi \{S(n-1) < 0 < S(n).\} \ldots (8)$$

Similarly to equation (6), the frequency of the signal is given by:

$$F\ SIGNAL(n) = ZCRn * Fs \qquad \ldots\ldots\ldots (9)$$

Finally, the frequency (*FSIGNAL*) tracked by the method is the arithmetic mean between both values found in (6) and (9), and it is calculated by the expression below:

F SIGNAL =(*F SIGNAL(n)* + *F SIGNAL(p))* **/ 2**.... (10)

B Autocorrelation (ACF)

Classical methodology: The autocorrelation function (ACF) is a sophisticated method largely used in Pitch Detection Algorithms (PDA). The discrete autocorrelation function Φ at lag τ for a discrete signal x(n) is defined by:

$\Phi(\tau)=1/N \sum X(n)\ X(n-\tau)$ ....(11)

For a pure tone, the ACF exhibits peaks at lags corresponding to the period and its integral multiples. The peak in the ACF of expression (11) at the lag corresponding to the signal period will be higher than that at the lag values corresponding to multiples of the period. For a musical tone consisting of the fundamental frequency component and several harmonics, one of the peaks due to each of the higher harmonics occurs at the same lag position as that corresponding to the fundamental, in addition to several other integer multiples of the period (subharmonics) of each harmonic.
Thus, a large peak corresponding to the sum contribution of all spectral components occurs at the period of the fundamental (and higher integral multiples of the period of the fundamental). This property of the ACF makes it very suitable for the pitch tracking of monophonic musical signals. The ACF PDA chooses as the pitch period the lag corresponding to the highest peak within a range of lags. Since the signals are commonly noisy, some extra processing is needed to increase the performance of the ACF. In this work some procedures were implemented.

Proposed methodology for musical notes: In order to increase the performance of the classical method, this work proposes the elimination of the low peaks in the ACF, leaving the highest peaks (the distance between these peaks is related to the period of the analyzed signal). Using these two peaks it is possible to detect the fundamental frequency with good accuracy. For the purpose of eliminating undesirable peaks of
the ACF, the vector Φ(|) is analyzed and its maximum is obtained as follows:

$\Phi max = MAX\ \{\Phi(n)\} = \Phi(nMAXFI)$ ..... (12)

The term *nMAXFI* is the index of this maximum From this index (that occurs always at the central point of the ACF vector), the first zero-crossing left is founded and its index is named *nCRUZ*. A new vector is created as follows:

$\alpha(n)=\Phi(n)$ 0<n<ncruz ..... (13)

Due to the characteristics of ACF, if the distance between the highest peak and the second highest peak (this is the highest one in the new vector $\alpha(n)$), the signal frequency can be determined. Thus, th maximum of $\alpha(n)$ is found as follows:

α MAX = max{ α(n)} = α (nMAXALFA) .....(14)

The position inside the vector containing this peak has the index *nMAXALFA*. A third vector *A(n)* is created, with the same values of $\Phi(n)$ but all values below *a(nMAXALFA)* are
eliminated (value 0 is assigned), what keep only the peaks used to detect the frequency without errors.

$$A(n) = \begin{cases} \Phi(n) & , \Phi(n) > \alpha(nMAXALFA) \\ 0 & , \Phi(n) < \alpha(nMAXALFA) \end{cases} \quad ....(15)$$

Since the number of samples (*nMAXFI* - *nMAXALFA*) is known, it is possible to determine the time between this two points (because the sampling frequency is known and fixed). This time is the same as the period of the fundamental frequency of the signal. Therefore, the frequency accepted as the pitch of the signal, detected by the ACF is obtained as follows:

$$F = (Fs) / nMAXFI - nMAXALFA \quad .....(16)$$

Note	Note (octave)	Frequency (Hz)	Note	Note (octave)	Frequency (Hz)
1	A(2)	109.2	10	D(2)	146.4
2	A(3)	219.4	11	D(3)	293.6
3	A(4)	440.3	12	E(1)	82.3
4	B(2)	123.6	13	E(2)	164.1
5	B(3)	247.4	14	E(3)	330.1
6	B(4)	494.8	15	F(1)	87.2
7	C(2)	130.8	16	F(2)	173.8
8	C(3)	261.3	17	F(3)	349.2
9	C(4)	524.3	18	G(1)	97.8

Table I. Frequencies used as reference

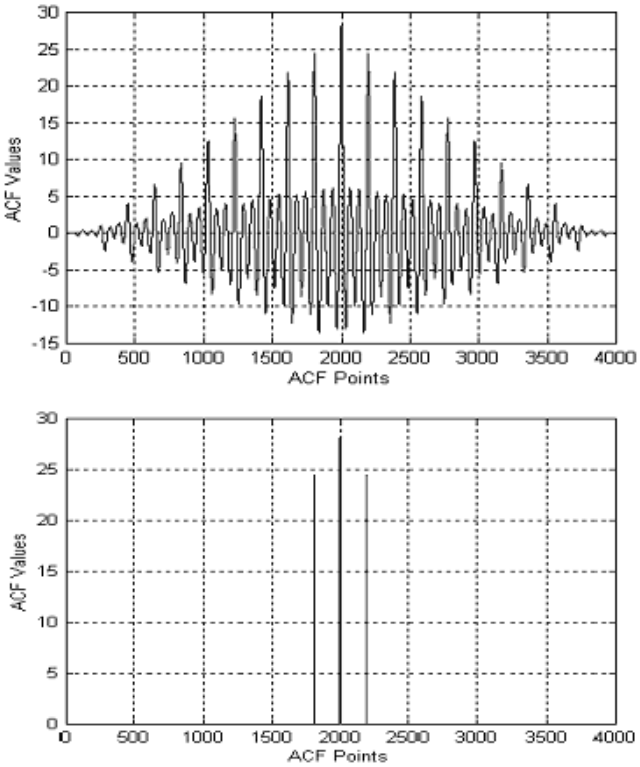


Fig. ACF in its classical approach applied to the note E(1) (upper) and with the methodology proposed applied (lower).

SEGMENTATION

Sounds generated by voice both in the form of speech and singing can be divided in voiced and unvoiced. Since in a performance of singing we are interested in music, the voiced regions need to be primarily investigated. However, the unvoiced segments must be retained for reconstructing the right duration of each event. Therefore, pitch tracking will be conducted only on voiced segments, while the information held by unvoiced regions will he employed in a final step. As it is illustrated in Fig., a first level of segmentation is achieved by comparing the RMS power of the signal to the threshold for signallnoise established in the previous section. Then, the derivative of the signal normalized to peak is used to emphasize the difference in amplitude. A procedure based on RMS power and zero crossing rate estimation is employed to take the voicedunvoiced decision. Events shorter or longer than a set of predefmed thresholds (e.g. consonants can not be longer than 260 msec) are removed or joined together. A further refinement in the segmentation procedure will be obtained through the pitch tracking stage in the case of notes sung 'legato'.

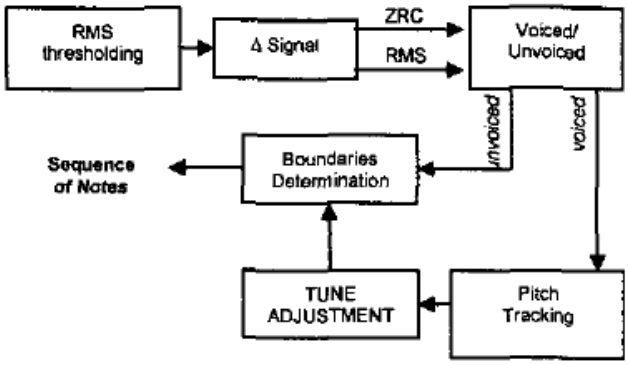


Fig. Block diagram of signal analysis from the segmentation stage to the estimation of sung notes.

REFERENCES

[1] Rafael George Amado Jozue Vieira Filho, *Pitch Detection Algorithms Based on Zero-Cross Rate and Autocorrelation Function for Musical Notes*

[2] MASRI *"Musical synthesis and modeling of voice"* .

[3] Stormy Attaway. *"A practical approach to matlab"*

[4] P.Ramesh babu *"Digital Signal processing"* .

[5] "*Jinachitra_Thesis a two side formanant analysis*".

[6] "*Music theory for Dummies"* & *"Computer Sound design "*

[7] "*Complete refrence Java*".

[8] John.G.Proakis and Dimitris G Manolakis,"*Digital Signal Processing,Principles,Algorithms,and Application*

IT 024

AUDIO SEGMENTATION

Name of Main Author :Mrs.Borawake Madhuri Pravin

Designation : Lecturer

Name of organization :College of Engineering ,Manajri,41207

Pune,411028 Maharashtra, India

madhuri.borawake@gmail.com,madhuri_borawake@yahoo.co.in

Contact : 9823353507,9823353524

Research Scholar form JJT,University

Name of Co-Author :Prof..Kawitkar Rameshwar

Designation : Professor

Name of organization :Sinhgad College Of Engineering, Pune

Pune, Maharashtra, India

rskawitkar@rediffmail.com

Contact : 9890551983

Name of Co-Author :Prof..Khadhatre Mahesh

Name of organization :Sinhgad College Of Engineering, Pune

Pune, Maharashtra, India

rskawitkar@rediffmail.com

ABSTRACT : This project describes the work done on the development of an audio segmentation and classification system. Many existing works on audio classification deal with the problem of classifying known homogeneous audio segments. In this work, audio recordings are divided into acoustically similar regions and classified into basic audio types such as speech, music or silence. Audio features used in this project include Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate and Short Term Energy (STE). These features were extracted from audio files that were stored in a WAV format. Possible use of features, which are extracted directly from MPEG audio files, is also considered. Statistical based methods are used to segment and classify audio signals using these features. The classification methods used include the General Mixture Model (GMM) and the k- Nearest Neighbour (*k*-NN) algorithms. It is shown that the system implemented achieves an accuracy rate of more than 95% for discrete audio classification.

Keywords: *audio content analysis, segmentation, classification, GMM, k-NN, MFCC, ZCR, STE and MPEG*

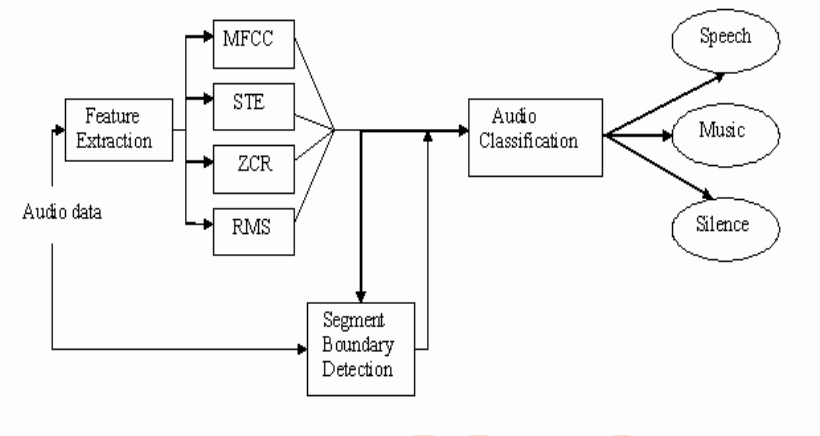
Introduction

Audio segmentation and classification have applications in wide areas. For instance, content based audio classification and retrieval is broadly used in the entertainment industry, audio archive management, commercial music usage, surveillance, etc. There are many digital audio databases on the World Wide Web nowadays; here audio segmentation and classification would be needed for audio searching and indexing. Recently, there has been a great deal of interest in monitoring broadcast news programs, in this case classification of speech data in terms of speaker could help in efficient navigation through broadcast news archives.

Like many other pattern classification tasks, audio classification is made up of two main sections: a signal processing section and a classification section. The signal processing part deals with the extraction of features from the audio signal. The various methods of time-frequency analysis developed for processing audio signals, in many cases originally developed for speech

processing, are used. The classification part deals with classifying data based on the statistical information extracted from the signals.

Two different classifiers, k-Nearest Neighbour(k-NN) and General Mixture model (GMM), were trained and tested to classify audio signals into music, speech and silence. The audio features used for classification were the Mel Frequency Cepestral Coefficients(MFCC), Zero Crossing Rates(ZCR) and Short Time Energy(STE). And for segmentation purposes Root Mean Square(RMS) features were used.



Segmentation and classification of audio data.

Audio feature extraction

Feature extraction is the process of converting an audio signal into a sequence of feature vectors carrying characteristic information about the signal. These vectors are used as basis for various types of audio analysis algorithms. It is typical for audio analysis algorithms to be based on features computed on a window basis. These window based features can be considered as short time description of the signal for that particular moment in time.

The performance of a set of features depends on the application. The design of descriptive features for a specific application is hence the main challenge in building audio classification systems. A wide range of audio

features exist for classification tasks. These features can be divided into two categories: time domain and frequency domain features. The Features considered in this chapter are: Mel Frequency Cepstral coefficient (MFCC), zero crossing rates and short time energy.

Audio Classification

The problem of classifying the extracted audio features into one of a number of audio classes is considered. The basic classification task can be considered as a process

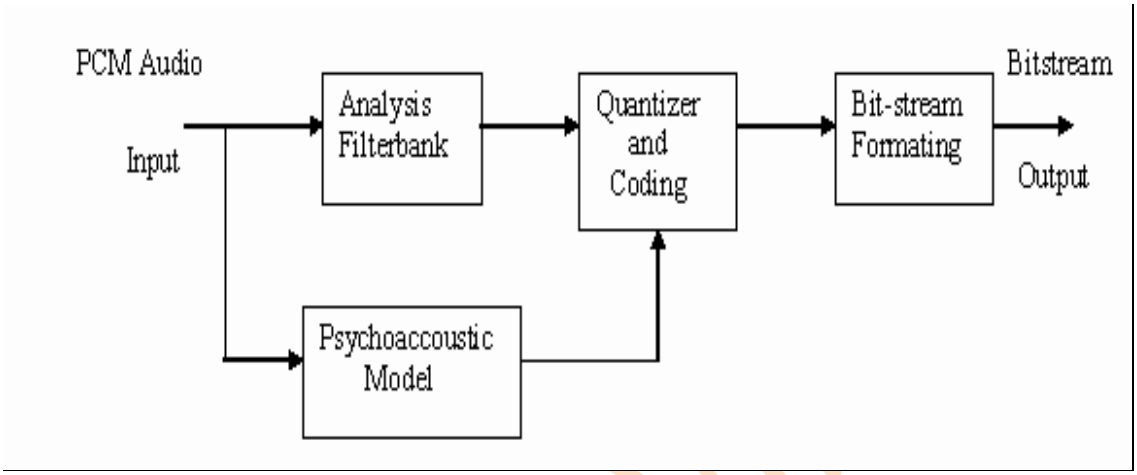
where a previously unknown input data is assigned to a class

$C \in \{C_1, C_2, ..., C_n\}$. Such assignments are made by establishing and applying a decision rule; for example, a simple decision rule could be the assignment of a new data sample to a class whose mean it is closest to in feature space.

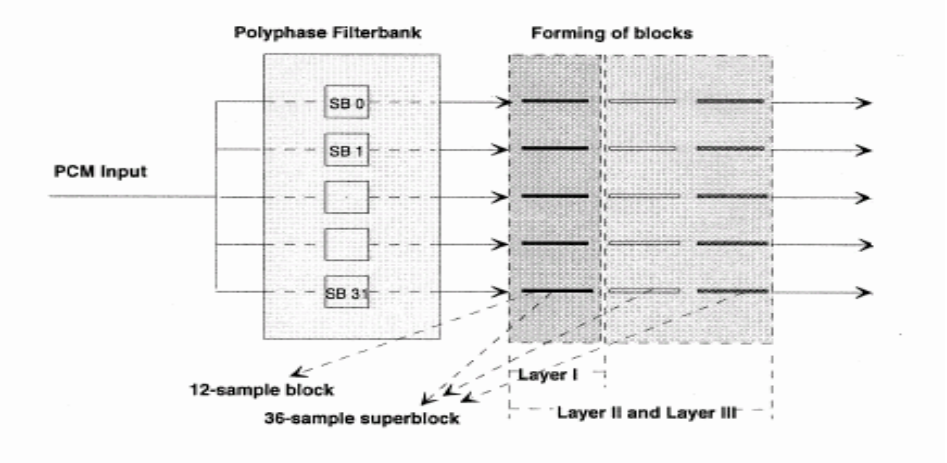
Classification algorithms are divided into supervised and unsupervised algorithms. In a supervised classification, a labelled set of training samples is used to "train" the algorithm whereas in the case of an unsupervised classification the data is grouped into some clusters without the use of labelled training set. Parametric and nonparametric classification is another way of categorizing classification algorithms. The functional form of the probably density of the feature vectors of each class is known in parametric methods. In non parametric methods on the other hand, no specific functional form is assumed in advance, instead the probability density is rather approximated locally based on the training data.

MPEG Audio Compression

In the following, a short description of the coding methods for the three MPEG-1 layers is given.



Block diagram of MPEG encoding



Subband blocks in MPEG encoding

Description of the audio data

The audio files used in the experiment were randomly collected from the internet and from the audio data base at IMM. The speech audio files were selected from both Danish and English language audios, and included both male and female speakers. The music audio samples were selected from various categories and consist of almost all musical genres. These files were in different formats (MP3, aif, wav, etc) and in order to have a common format for all the audio files and to be able to use them in matlab programs, it was necessary to convert these files to a wav format with a common sampling frequency. For this purpose the windows audio recorder was used and the recorded audio files were finally stored as 22050 Hz, 8 bit, mono audio files. The recorded audio files were further partitioned into two parts: the training set and the test set. This was important since each audio file was intended to be used only once, either for training or for testing a classifier. The training vectors correspond to 52566 frames for speech and 73831frames for music files.

Audio type	Number of files	Average length	Total length
Speech	45	15 sec.	675 sec.
Music	55	15 sec.	825 sec.

Training data

Audio type	Number of files	Average length	Total length
Speech	30	15 sec.	450 sec.
Music	30	15 sec.	450 sec.
silence	30	15 sec.	450 sec.
Music +	10	120 sec.	1200 sec.

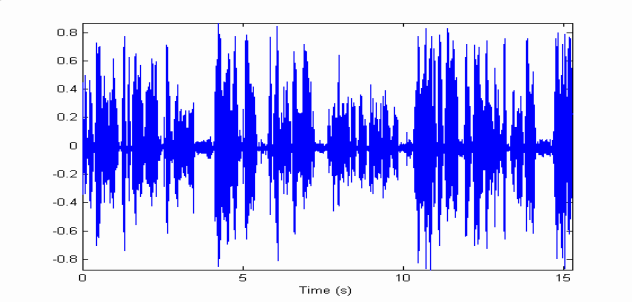
Test data

MFCC features

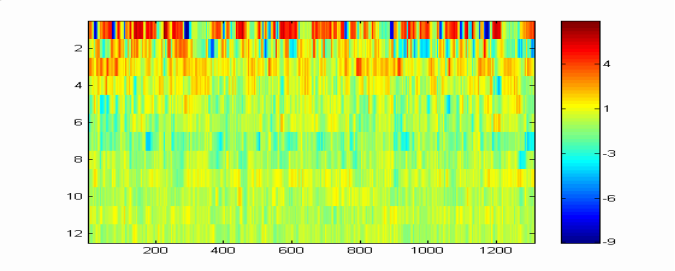
In order to extract MFCC features from the row audio signal, the signal was first partitioned into short overlapping frames each consisting of 512 samples. The overlap size was set to half the size of the frame. A Hamming window was

then used to window each frame to avoid signal discontinuities at the beginning and end of each frame. A time series of MFCC vectors are then computed by iterating over the audio file resulting in thirteen coefficients per frame. The actual features used for classification task were the means of the MFCCs taken over a window containing 15 frames. Furthermore only six out of the thirteen coefficients were used. In this way a very compact data set was created. The following figures show plots of the speech and music signals as a function of time together with

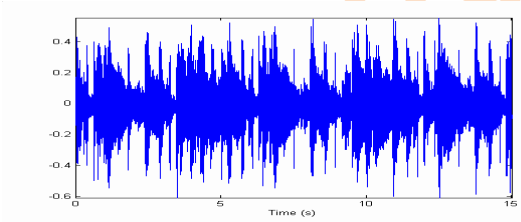
their respective MFCCs .



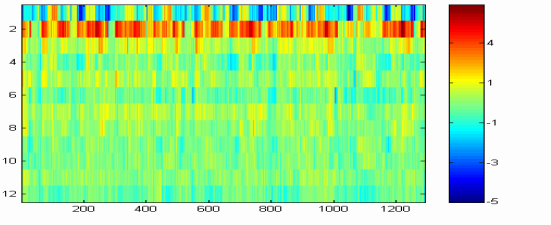
Plot of the MFCCs for the speech signal function of time



Plot of a speech signal as



Plot of a music signal as function of time signal



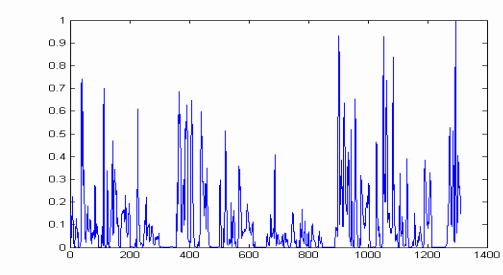
Plot of the MFCCs for the music

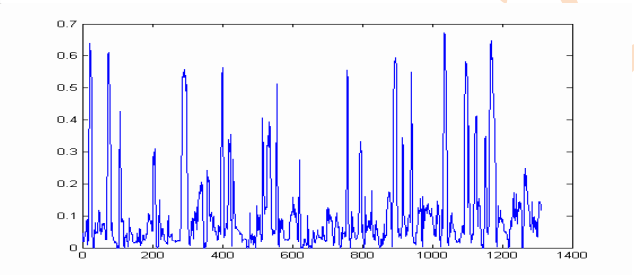
The STE and ZCR features

Since these features were intended to be usedeither in conjunction with the MFCCs or independently, it was necessary to split the audio signal so that the

[CONFERENCE PROCEEDINGS ISBN 978-81-921445-1-1
COLLECTION OF PAPERS SUBMITED FOR CONFERENCE]

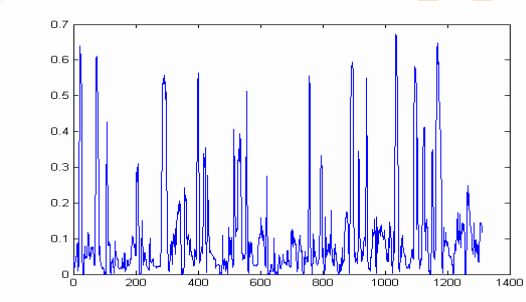
length of these features were the same as the length of the MFCCs. Hence, the partition of the audio signal into overlapping windows was exactly the same as in the case of the MFCC features. The Short Time Energies and the Zero-crossing rates were extracted from such windows, one from each window. The actual features used for the classification task were the means taken over a window containing 15 frames. The following figures show plots of STE and ZCR for both music and speech signals.

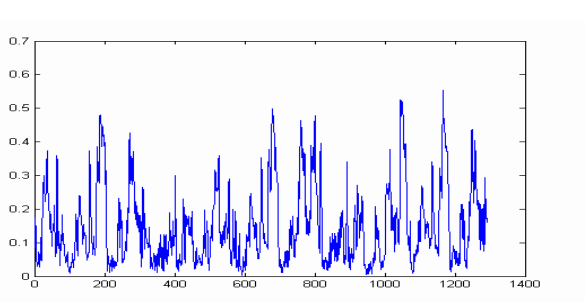




STE for speech signal

STE for music signal



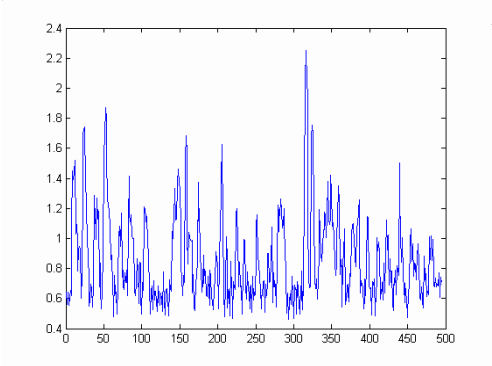


ZCR for speech signal

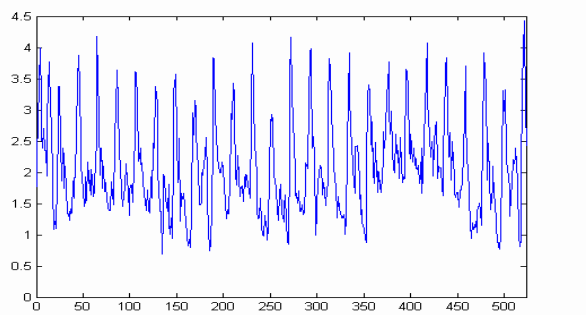
ZCR for music signal

The RMS feature

Although the RMS is somewhat related to the short time energy, it is often used as a measure of the loudness of audio signals and therefore a unique feature to segmentation. Since this feature was used alone for another task, the audio signal was split in a rather different way. The audio signal was first partitioned into short non overlapping frames each consisting of 512 samples. The Root Mean Square was computed by iterating over the audio file based on the amplitude equation shown on page 25 and a single RMS value is obtained for each frame. The following figures show plots of RMS for both music and speech signals.



RMS of a speech signal



RMS of a music signal

	Music	**Speech**
Music	**89.1**	10.9
Speech	8.4	**91.6**

Confusion matrix when MFCC features were the only inputs.

	Music	**Speech**

Music	**93.01**	6.99
Speech	5.49	**94.51**

Confusion matrix when the inputs were MFCC and STE features.

	Music	**Speech**
Music	**90.6**	9.39
Speech	7.4	**92.53**

Confusion matrix when the inputs were the MFCC and ZCR.

	Music	**Speech**
Music	**93.6**	6.33
Speech	5.7	**94.29**

Confusion matrix when the inputs were MFCC, ZCR and STE features. From the results obtained the following observations can be made. The MFCC features used as an input, alone, result in an overall correct classification rate of 90.3%. When the MFCC features were used in conjunction with the short time energy and the Zero Crossing rate the overall classification rate gets better and is around 93.98%. The same is true when MFCC feature are used together with short time energy features. However, when the input to the classifier was a combination of MFCC features and zero crossing

rate only little improvement in the overall correct classification rate was

seen. We conclude therefore that the MFCC features in conjunction with the short time energy alone can with a good classification rate be used for a speech/music discrimination.

	Music	Speech
Music	**85.22**	**14.78**
Speech	**0.4**	**99.56**

The features used were the MFCCs.

	Music	Speech
Music	**89.78**	**10.22**
Speech	**0.22**	**99.78**

The features used were the MFCC and STE features.

	Music	Speech
Music	**85.65**	**14.35**
Speech	**0.00**	**100.00**

The features used were the MFCC and ZCR features.

	Music	Speech

Music	**91.30**	**8.70**
Speech	**0.00**	**100.00**

The features used were the MFCC, STE and ZCR features.

Although the results obtained in this case showed similar tendencies as in the case of the K-nearest neighbour classifier, the correct classification rate was even better. When the MFCC features were used in conjunction with the short time energy and zero crossing rate, a correct classification rate of around 95.65% was obtained. This result was the best result among the classification results obtained from both the GMM classifier and the KNN classifiers. A correct classification rate of about 94.78% was obtained for the case when MFCC in conjunction with the Short Time Energy features were used. However, for the case where the input was a combination of MFCC and ZCR features, the classification rate was 92.83% , which is almost the same as when pure MFCC features were used.

Comparison of the classification results

The Table below shows the classification results obtained for the two classifiers with different feature combinations.

Features	**k-NN (k=5) Accuracy (%)**	**GMM Accuracy (%)**
MFCC	90.35	92.39
MFCC + ZCR	91.57	92.83
MFCC + STE	93.76	94.78

MFCC + ZCR+ STE	93.98	95.65

Accuracy testing results for the speech/music classifier

The classification results obtained from using a general mixture model classifier and a k- nearest neighbour classifier demonstrate the effect of the classification algorithms. The general mixture model classifier seemed to have a far better correct classification rate than k-nearest neighbour classifier (around 2%). In both cases, adding more features to the MFFC features showed a positive effect on the outcome, although using the MFCC in conjunction with STE resulted in a relatively higher classification rate than when MFCC features were used in conjunction with the zero crossing rates.

Conclusion and future work

The aim of this project was to design a system that could be used to segment an audio signal into similar regions and then classify these regions into music, speech and silence audio classes. The project could be considered as a combination of two tasks; a segmentation task and a classification task. Classification algorithms were used either independently with a given audio segment or in combination with the segmentation algorithm.

Features extracted from music and speech signals (in WAV format) were used in the two tasks. Three feature sets were used to train and test two different classifiers, the General Mixture Model classifier and the k-Nearest Neighbour classifiers, to classify audio signals, and only one feature set was used to partition audio into similar regions. Nearly all the audio files used in this project had been obtained from the internet. The majority of these audio files were in MP3 format and it was necessary to convert them to WAV format. Thus, the process for extracting audio feature showed to be very time consuming. It would have been very advantageous if the system was designed to take in audio in MP3 format. This could have had two effects on the system; the need for converting one audio format to another would have been avoided, and features would have been extracted directly from the encoded data. The two classifiers were trained and tested with the same training and test sets. With

each classifier, four experiments were run with different combinations of the feature sets.

The system implemented worked well on classifying any type of music and speech segments with a correct classification rate of 95.65% for one second windows. The system also worked reasonably well for segmenting audio signals into similar classes. Some improvement in the segmentation method used is however required.

There are many things that could be done in the future. The segmentation algorithm could be modified to detect the transition point with an accuracy of 30ms, and also to automatically set the threshold for finding the local maxima of the normalised distance measure. More training data could be used in the classification part. The system could be trained to include other classes other than music, speech and silence. Further classifications into different music genre or identifying a speaker are also other possibilities.

References

[1] Lie Lu, Hong-Jiang Zhang and Hao Jiang. "Content analysis for audio classification and segmentation". *IEEE Transactions on speech and audio processing,* vol.10, no.7, October 2002

[2] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal , " Speech/Music discrimination for multimedia applications," *Proc. IEEE Int. Conf. on acoustics, Speech, Signal Processing* (Istanbul), pp. 2445-2448, June 2000

[3] H. Meindo and J.Neto, " Audio Segmentaion, Classification and Clustering in a Broadcast News Task" , *in Proceedings ICASSP 2003*, Hong Kong, China, 2003.

[4] G. Tzanetakis and P. Cook, " Multifeature audio segmentation for browsing and annotation," *Proc.1999 IEEE workshop on applications of signal processing to Audio and Acoustics*, New Paltz, New York, Oct17-20, 1999.

[5] C. Panagiotakis and G.Tziritas " A Speech/Music Discriminator Based on RMS and Zero-Crossings". *IEEE Transactions on multimedia,* 2004.

[6] E. Scheirer and M. Slaney, " Construction and evaluation of a robust

multifeature speech/music discriminator, " in *Proc. ICASSP '97*, Munich, Germany, 1997, , pp. 1331-1334.

[7] Davis Pan, "A Tutorial on MPEG/Audio Compression,". IEEE Multimedia Vol. 2, No. 7, 1995, pp. 60-74.

[8] Silvia Pfeiffer and Thomas Vincent "Formalisation of MPEG-1 compressed domain audio features", Technical Report No.01/196, CSIRO Mathematical and Information Sciences, Dec. 2001.

[9] G. Tzanetakis and P. Cook, " Sound analysis using MPEG compressed audio", *Proc.*

IEEE Intl. Conf. on acoustics, Speech, Signal Processing, ICASSP, 2000

[10] D. Pan, " A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, No.2, 1995, pp.60-74.

[11] Christopher M. Bishop, *Neural Networks for Pattern Recognition* , Oxford

University Press, 1995

[12] Tong Zhang and C.C. Jay Kuo, "Heuristic Approach for Generic Audio Data

Segmentation and Annotation," *ACM Multimedia (1),* 1999, pp 67-76.

[13] Beth Logan, " Mel Frequency Cepstral Coeffcients for Music Modelling," *in international Symposium on Music information retrieval*, October 2000.

[14] John R. Deller, Jr., John H.L. Hansen and John G. Proakis, *Discrete-Time Processing of Speech Signals,* IEEE Inc. 2000.

[15] John G. Proakis and Dimitris G. Manolakis, *Digital Signal Processing principles, algorithms and applications,* Prentice-Hall, Inc, 1996.

[16] L.R. Rabiner and R.W.Schafer, *Digital Processing of speech signals*, Prentice-Hall, 1978.

[17] MPEG Maaate. http://www.cmis.csiro.au/Maaate/

IT 025

LINGUISTIC ANALYSIS TO DETECT AMBIGUITIES AND INACCURACIES IN USE-CASES

Saket Guntoorkar, Rohan Kokandakar

Saiprasad Dhumal, Yogesh Arvi

Students B.E. Computer Engineering

PVG's COET, University Of Pune, India

saket.guntoorkar@gmail.com, rohanhk8@gmail.com

Abstract—

Numerous tools and techniques are available for managing requirements. Many are designed to define requirements, provide configuration management, and control distribution. However, there are few automatic tools to support the quality analysis of natural language (NL) requirements. Ambiguity analysis and consistency and completeness verification are usually carried out by human reviewers who read requirements documents and look for defects. This clerical activity is boring, time consuming, and often ineffective. This report describes a disciplined method and designing of an automated tool that can be used for the analysis of NL requirement documents.

System behaviour cannot be specified in detail with Use Case diagrams. The technique is based on natural language specification for scenarios and extensions. Scenarios and extensions are specified by phrases in plain English language. This makes requirements documents easy to understand and communicate even to non-technical people. Natural language is powerful (the expression power of the English language is said to be higher than any other language in the world), well known and generally easy to understand.

Every project or a work requires a set of instructions or data using which one can proceed onto his/her work. The data or set of instructions must be correct in terms of its spelling and meaning. The data must be particular not vague. When expressing goals, scenarios and conditions with NL, it is necessary to identify the defects due to the inherent ambiguity of NL (for instance: vagueness, poor

specification and poor accuracy). For this reason, tools and techniques that were conceived for traditional textual requirements can be effectively applied to Use Cases to detect defects and collect metrics. We will be using in built parsers like shallow parser, deep parser, clang parser, mini parser. Also we will use tools like dictionary concept annotator, anaphora resolver to convert pronouns into nouns, context annotator, process builder, expression builder. Project work is to develop an algorithm from existing libraries, parsers and lexers to detect ambiguities, inaccuracies in use case to get high quality and meaningful information from it.

Keywords: Natural language, Linguistic, Use-case, Ambiguities, Parser, Metrics

1.Introduction

It is well known that defects in software requirements have an adverse impact on dependability. Formal methods of requirement specification are often suggested as a remedy for errors in requirement elicitation. In spite of the multiple benefits that formal methods have to offer to a requirements analyst, their adoption has been slow in software engineering processes (except possibly in the fields of hardware design and safety critical control systems). This is primarily due to the fact that the adoption overhead (learning, tool support, configurability, etc.) still overshadows the economic gain.

Also, with the onset of agile methods, the emphasis is on active and continuous participation of the customer in the development process. Use of formal methods for requirement elicitation introduce high entry barrier for customer participation. Further, most formal notations are geared towards expressing and validating specific kind (functional, performance, safety, security, etc.) of requirements, which often make them inappropriate for a general purpose requirement document.

The tool will be consisting of Lexical analyzer as well as Syntactic analyzer. This tool will reduce the interpretation problem of NL statements in use-cases or in business cases. The quality metrics can also be derived from this tool. The tool along with the human interaction will help in reducing misinterpretation of NL.

In this paper, we limit our discussions to the analysis of textual use case requirements.

In particular, in this paper we present a novel and domain independent linguistic technique. To the best of our knowledge, this is the first report on effectiveness of applying linguistic analysis to large and diverse set of industrial use cases.

2. Use cases

A *Use Case* describes the interaction (triggered by an external actor in order to achieve a goal) between a system and its environment. A *Use Case* defines a *goal*-oriented set of interactions between external actors and the system under consideration. The term *actor* is used to describe the person or system that has a goal against the system under discussion. A primary actor triggers the system behaviour in order to achieve a certain goal. A secondary actor interacts with the system but does not trigger the Use Case.

A Use Case is completed successfully when that goal is satisfied. Use Case descriptions also include possible extensions to this sequence, e.g., alternative sequences that may also satisfy the goal, as well as sequences that may lead to failure in completing the service in case of exceptional behaviour, error handling.

A complete set of Use Cases specifies all the different ways to use the system, and therefore defines the whole required behaviour of the system. Generally, Use Case steps are written in an easy-to-understand, structured narrative using the vocabulary of the domain. The language used for the description is English. Any other natural language can be used as well, and although our analysis focuses on English, the same reasoning can be applied to other languages (considering the obvious differences in semantics). The usage of natural language is engaging for final users who can easily follow and validate the Use Cases, and encourages them to be actively involved in defining the requirements. A scenario is an instance of a Use Case, and represents a single path through the Use Case. Thus, there exist a scenario for the main flow through the Use Case, and other scenarios for each possible variation of flow through the Use Case (e.g., triggered by options, error conditions, security breaches, etc.).

An example of Use-case for a car insurance company

Primary Actor: the claimant

Goal: Get paid for car accident

Scope: The Insurance Company (MyInsCo)

Level: Summary

Stakeholders and Interests:

The claimant – to get paid the most possible

MyInsCo – to pay the smallest appropriate amount

The dept. of insurance – to see that all guidelines are followed

Precondition: none

Minimal guarantees: MyInsCo logs the claim and all activities

Success guarantees: Claimant and MyInsCo agree on amount to be paid, claimant gets paid that.

Trigger: Claimant submits a claim

Main success scenario:

1. Claimant submits claim with substantiating data.

2. Insurance company verifies claimant owns a valid policy

3. Insurance company assigns agent to examine case

4. Agent verifies all details are within policy guidelines

5. Insurance company pays claimant

Extensions:

1a. Submitted data is incomplete:

1a1. Insurance Company requests missing information

1a2. Claimant supplies missing information

2a. Claimant does not own a valid policy:

2a1. Insurance company declines claim, notifies claimant, records all this, terminates proceedings.

3a. No agents are available at this time

3a1. (What does the insurance company do here?)

4a. Accident violates basic policy guidelines:

4a1. Insurance company declines claim, notifies claimant, records all this, terminates proceedings.

4b. Accident violates some minor policy guidelines:

4b1. Insurance Company begins negotiation with claimant as to degree of payment to be made.

3. Linguistic Analysis and its need

NL plays a relevant role in the specification of requirements by Use Cases because actors, actions, scenarios, responsibilities, goal etc. are specified in NL. The use of NL as a way to specify the behaviour of a system is always a critical point, due to the inherent ambiguity originating from different interpretations of natural language descriptions. Instead of NL, UML (Unified Modeling Language) modeling tools can be used for diagrammatic representation of requirement documents. But in this case, only technical people related to project can understand the requirements while it will be difficult for a common man to understand.

Linguistic analysis means scientific study of the natural language we speak. The business cases or simply the use-cases written in natural language must be processed in order to detect its ambiguities. The tool which we are developing helps in detecting those ambiguities. During linguistic analysis the input text goes through following stages: Lexical analysis, syntax analysis, semantic analysis, ambiguity detection, ambiguity correction by user and reiteration of previous stages (if user wants) till complete removal of ambiguities.

4. Ambiguities in use-cases

The analysis made by means of NL-based techniques is useful to address several interpretation problems related to linguistic aspects of Use Cases. These problems may be grouped into three main categories:

- *Expressiveness* category: it includes those characteristics dealing with the understanding of the meaning of Use Cases by humans. The following topics have to be considered as part of the Expressiveness category:
1. *Ambiguity mitigation*: detection and correction of linguistic ambiguities in the sentences;

2. *Understandability improvement*: evaluation of the understandability level of a requirements specification document and indication of the parts of it needing to be improved;

Consistency category: it includes those characteristics dealing with the presence of semantics contradictions and structural incongruities in the NL requirements document.
Completeness category: it includes those characteristics dealing with the lack of necessary parts within the requirements specifications.eg incomplete template of a use-case.

Table1 shows expressiveness defect indicator.

Expressiveness	Vagueness	words having a non uniquely quantifiable meaning
	Subjectivity	sentence refers to personal opinions or feeling
	Weakness	sentence contains a weak main verb
	Under Specification	when the subject of the sentence contains a word identifying a class of objects without a modifier specifying an instance of this class
Understandability	Multiplicity	if the sentence has more than one main verb
	Unexplanation	when a sentence contain acronyms not explicitly and completely explained
Consistency	Under Reference	when a sentence contains explicit references to: not numbered sentences, documents not referenced into and entities not defined nor described

Table1 Expressiveness Indicators

5. How this tool will work?

The tool will take input use-cases in text format. The input will first go through lexical analysis phase. The lexical processor tokenizes the text into words and/or punctuation marks, determines the base form of a word and associates words with contextually appropriate part-of speech (POS) information.

The next step is to pass the output of lexical phase to shallow parser. It acts as a syntactic analysis system for identification of phrasal, configurational and grammatical information in free text documents. The shallow parser can operate just on the basis of the tokens and POS information provided by the lexical processor. This phase makes a parse tree of the input string and will find out (or highlight) ambiguous words from the use-case. The user will replace those words with suitable and unambiguous words and can again check the use-case in same respect.

To check the consistency, the use-case must be semantically analyzed. This step is carried out by Semantic analysis phase. This stage will derive the meaning of a complete sentence and also its relation with the previous sentence. This will help in finding out the consistency of statements in use-case.

There are certain cases where the user might feel that a word shown ambiguous by our tool won't be ambiguous then in such case, user can unmark that word. E.g. an ambiguous word in one domain might not be ambiguous in other. In this way the tool will highlight the ambiguities and the user may make changes by replacing the ambiguous sentences of his/her use-case.

6. Conclusion and Future work

We have demonstrated in the paper that our approach is able to extract and signify the ambiguities in a use-case. We have proposed the use of easily available linguistic techniques to support semantic analysis of Use Cases. Linguistic techniques may provide an effective support towards the achievement of quality requirements by extracting useful information from them but are not sufficient to completely address the aspects of correctness and consistency check of requirements. Using this tool performance matrix can be derived to find effectiveness of tool for analysing given text input and quality information it depicts after analysis. The application of linguistic techniques is promising and prefigures a very cost-effective support for developers if carried out at early

stages of the requirement process. The tool along with a person handling will reduce the ambiguities to a greater extent. Future work is to make the tool platform independent, input file type independent and domain independent by using better parsing techniques.

References

[1] Avik Sinha, Amit Paradkar, Palani Kumanan, Branimir Boguraev

"A Linguistic Analysis Engine for Natural Language Use Case Description and Its Application to Dependability Analysis in Industrial Use Cases". 2009

[2] A.FANTECHI, S.GNESI, G.LAMI, A. MACCARI "Application of Linguistic Techniques for Use Case Analysis" 2002

[3] Systematic Web Data Mining with Business Architecture to Enhance Business Assessment Services, 978-0-7695-4371-0/11 © IEEE

[4] Improving the Efficiency of Legal E-Discovery Services using Text Mining Techniques. 2011

[5] D. Ferrucci and A. Lally, "UIMA: an architectural approach to unstructured information processing in the corporate research environment," Natural Language Engineering, vol. 10, no. 4, 2004.

[6] Open Source Software and Coding Information site "www.sourceforge.com".

[7] "www.wikipedia.com" for general information on different topics in project.

[8] Wilson, "Automated analysis of Requirement Specifications"

IT 026

An External Storage Support for Mobile Application with Scarce Resources

Pratima.J.Shedge,Nishad Rajwade

Computer department,University of Pune

College of Engineering Manjari(bk) Pune-411028

Pratimashedge8@gmail.com,

rajwadenishad@ymail.com

Abstract

Nowadays, users of mobile phones generate too many files that have to be frequently downloaded to an external storage repository, restricting the user mobility. This paper presents a File Transfer Service (FTS) for mobile phones with scare storage resources. It is a support that can be used through a set of functions (API) that facilitates file transfer between mobile applications and external storage servers, taking advantage of different wireless networks. The FTS selects the best wireless connection (WiFi, GPRS or UMTS) considering accessibility and cost of the service. FTS is able to use the Multimedia Messaging Service (MMS) as an alternative option for transferring files, which is especially useful when the mobile phone connectivity is limited. It is based on the J2ME platform. As a use case, a mobile application named Swapper was built on top of the FTS. When the mobile phone memory runs out, Swapper automatically sends selected files to a web storage server using the best connection available, increasing the storage space in the mobile phone. Swapper includes an efficient replacement policy that minimizes the latency perceived by users.

Keywords—Wi-Fi, GPRS, MMS, FTP

I. INTRODUCTION

The cell phone is one of the most popular mobile devices. It has constantly evolved since its invention. It's common that current mobile devices have two or more communication interfaces such as USB, infrared, Bluetooth, GPRS and the IEEE 802.11, popularly known as Wi-Fi. Wireless networks improve the utility of portable computing devices, enabling mobile user's versatile communication and continuous access to services and resources of the terrestrial networks. Mobile phones have also been endowed with more computing power. Intelligent mobile phones (a.k.a. smartphones) integrate the functions of PDA (Personal Digital Assistant) in conventional mobile phones. These technologic developments allow current mobile phones to run applications that generate a large number of files and, as a consequence, require a greater storage capacity. Unfortunately, the storage capacity on these devices has not grown at the pace that mobile applications require. Once a mobile device exceeds its storage capacity, it is necessary to download its files (such as text, images, music,videos, etc.) in an external storage system (e.g., a PC), backing up the information. This process usually limits the mobility because users have to connect their mobile devices to a fixed external storage using a cable (e.g., USB), or through a short- range wireless network such as Infrared or Bluetooth. These storage limitations reduce the mobility and storage capacity of users, particularly in situations when users are travelling and they need more space to keep, for instance, their pictures or videos and there is not a nearby external storage device available. These situations motivate the development of supporting tools that allow files to be exchanged between mobile devices and external storage systems in a transparent way, taking into account LAN or WAN wireless connectivity, cost of the service and available storage. Since there are different mobile devices that work with various computing platforms or operating systems (e.g., Symbian, Blackberry, Windows Mobile, Android, etc.), it is not feasible to develop a proprietary tool for a specific platform, because it would limit its usability. That is why the design of a solution to this problem must include multiplatform considerations that allow better coverage in different mobile devices.

The situations described above were our motivation to build a File Transfer Service (FTS) that can be capable of running on different operating systems and allow mobile applications to send and receive files from an external storage server, connected to Internet through different wireless technologies, considering the cost of the service. The FTS was tested by means of a mobile application called Swapper, which offers a service of swapping files between

the mobile device and an external storage server in a transparent way. Users of this application perceive a virtual storage space, which is higher than the real memory space included in the mobile device. The swapping service is similar to a file caching service, reason why Swapper integrates an efficient replacing policy to optimize the file access time.

.The ability to access information on demand at any location confers competitive advantage on individuals in an increasingly mobile world. As users become more dependent on this ability, the span of access of data repositories will have to grow. The increasing social acceptance of the home or any other location as a place of work is a further impetus to the development of mechanisms for mobile information access.These considerations imply that data from shared file systems, relational databases, object-oriented databases, and other repositories must be accessible to programs running on mobile computers.

Coda is designed for an environment consisting of a large collection of untrusted Unix 1 clients and a much smaller number of trusted Unix file servers. The design is optimized for the access and sharing patterns typical of

academic and research environments. It is specifically not intended for

applications that exhibit highly concurrent, fine granularity data access.

Mobile file systems such as Coda[1, 2], Odyssey[3], Bayou [4] and Xmiddle [5], worked with the data sharing-oriented problem in distributed computing environments. This problem could be directly related to the file transfer problem in mobile phones. Although with different focus, all of these systems try to maximise the availability of the data using data replication, each one differing in the way that they maintain consistency in the replicas. Coda provides server replications and disconnected operations; it allows access of data during the disconnection period and focuses on long-term disconnections, which more often occurs in mobile computing environments. Odyssey is the successor of Coda, which has been improved introducing context-awareness and application-dependent behaviors that allow the use of these approaches in mobile computing settings.

The Bayou system is a platform to build collaborative applications. Its emphasis is on supporting application-specific conflict detection and resolution. It has been designed as a system to support data sharing among mobile users and is intended to run in mobile computing environments. The system uses a read-any/write-any replication scheme, thus the replicated data are only weakly consistent. Unlike previous systems,Bayou exploits application knowledge for dependency checks and merges procedures. Lui et al [6] propose a mobile file system, NFS/M,

based on the NFS 2.0 and the Linux platform. It supports client-side data caching in order to improve the system performance, reducing the latency during weak connection periods. Atkin and Birman [7] propose other file system that, like NFS/M, supports client- side caching. Some applications like [8, 9], enable the file transfer between mobile devices and external storage servers. However, these applications only consider a proprietary storage server.

In [10], Boulkenafed and Issarny present a middleware service that allows collaborative data sharing among ad hoc groups that are dynamically formed according to the connectivity achieved by the ad hoc WLAN. This middleware enable to share and manipulate common data in a collaborative manner (e.g, working meet, network gaming, etc.) without the need for any established infrastructure. They implemented their middleware service within a file system in order to evaluate it. The result was a distributed file system for mobile ad hoc data sharing. It is worth mentioning that the performance measurements were done on a platform of ten laptops, and they only use IEEE 802.11b WLAN in ad hoc mode, unlike FTS, which is able to use Wi-Fi, GSM, GPRS or UMTS networks. Belimpasakis et al [11] propose a content sharing middleware for mobile devices using different protocols (UPnP, Atom and WebDAV), providing interfaces for applications, in order to allow 3rd party developers to create applications with sharing capabilities. The middlewares mentioned above make use of both Wi-Fi and GPRS wireless networks. However, they consider neither transferring files through a messaging system like MMS nor the portability issue. We have decided to develop FTS using J2ME, offering portability.

III. THE FILE TRANSFER SERVICE ARCHITECTURE

In this section, the FTS architecture is described in more detail. The architecture is mainly divided into three components or layers: Client-side application layer, Core connection layer and server-side application layer. Figure 1 depicts this architecture whose components are presented in the following sub-sections.

A. Client- side application layer (API-ESS)

This layer deals with file transferring operations (send/receive) required from mobile applications. One main function is to make transparent the selection of the available wireless connection (WiFi, GPRS or UMTS) or messaging service (MMS) provided by the lower layer (Core) to the mobile application. The main component of the client-side application layer is the External Storage Services Module (API_ESS). This module offers a set of wrappers for connections with different wireless networks. It uses a connection selector that indicates which wireless network will be used. When

connectivity is limited, the connection selector offers the MMS service as the best option for file transferring.

The API ESS includes both messaging services SMS and MMS. Due to the limitations of SMS (about 150 bytes by message), MMS is the default option chosen by the connection selector.

The API_ESS could include a configuration file that indicates the priorities assigned to the wireless connection services. The first in the list indicates the cheapest service in terms of cost or bandwidth consumption. The API_ESS functions are divided into upper level functions and lower level functions. The upper functions are typically used by mobile applications, e.g., selectRoot( ), retr(file) and stor(file). These functions send and receive files from an external storage service making totally transparent the type of wireless o messaging service used. Lower level functions deal directly with wireless connection and messaging wrappers. Examples of these functions are: autoDetectConnection(), getConnection() and setConnection().

The current implementation of the API_ESS includes the following wrappers:

- Wi-Fi/GPRS: It enables to send/request files via the IEEE

802.11 and the GPRS protocols.

- MMS: It represents an alternative option for sending and requesting files using the Multimedia Messaging Service (MMS). MMS has to be supported by a mobile phone carrier; otherwise it will not be available. This wrapper will commonly be used when the Wi-Fi and GPRS (or UMTS) networks are not available.

- SMS-M: The purpose of this wrapper is to provide

mobile applications with one more option for sending information to and/or receiving from the external storage server in future applications. The mobile application Swapper uses this wrapper for registering user accounts in a web storage server that is supported by the server- side layer of the FTS.

B. Core connection layer

This part of the architecture represents the communication services needed by our service. It includes modules to deal with the most popular wireless technologies and messaging services. A brief description of these modules is as

follows:

- SMS Center (SMSC): The SMSC is responsible for relaying messages between short message entities (SMEs, elements that can send or receive text messages) and store and forward messages, if the recipient SME is not available

- MMS Center (MMSC) [1]: The MMSC is a key element in the MMS architecture, and is composed of an MMS relay and an MMS server, which are responsible both to store and manage incoming and outgoing multimedia messages, and ensure interoperability with other messaging systems [12] by means of different communication interfaces (e.g., MM3 interface).

- SMTP Server: It receives the files that have been sent via multimedia messaging (MMS) to an email account. The data travels across the wireless network available (2.5G or 3G) and are routed by the MMSC to the Internet, then get to our SMTP server. This option is activated when socket connections fail.

- GSM Gateway: It receives text messages (SMS) that are sent by the client for complementary service e.g.,to create an account.

- User DB: It is an optional database that could contain information of subscribers registered by mobile applications based on FTS.

- FTP Server: It is the process responsible for receiving and storing the files. In the Swapper application, usershave to be registered before obtaining a storage space in the external storage server. Files managed by the FTP Server can come from direct socket connected clients (using a wireless connection) or e-mail clients (using a MMS connection). This server is one of the key parts of the architecture as it controls all the files received in the external storage server.

- e-mail Client: This process is used by the server-side layer when the original server-side MMS receiver fails. In this situation, files that are sent by clients using MMS are redirected to a mail box defined in the server-side application layer. The e-mail Client process is in charge of obtaining these files from the indicated mail box using the Post Office Protocol (POP3).

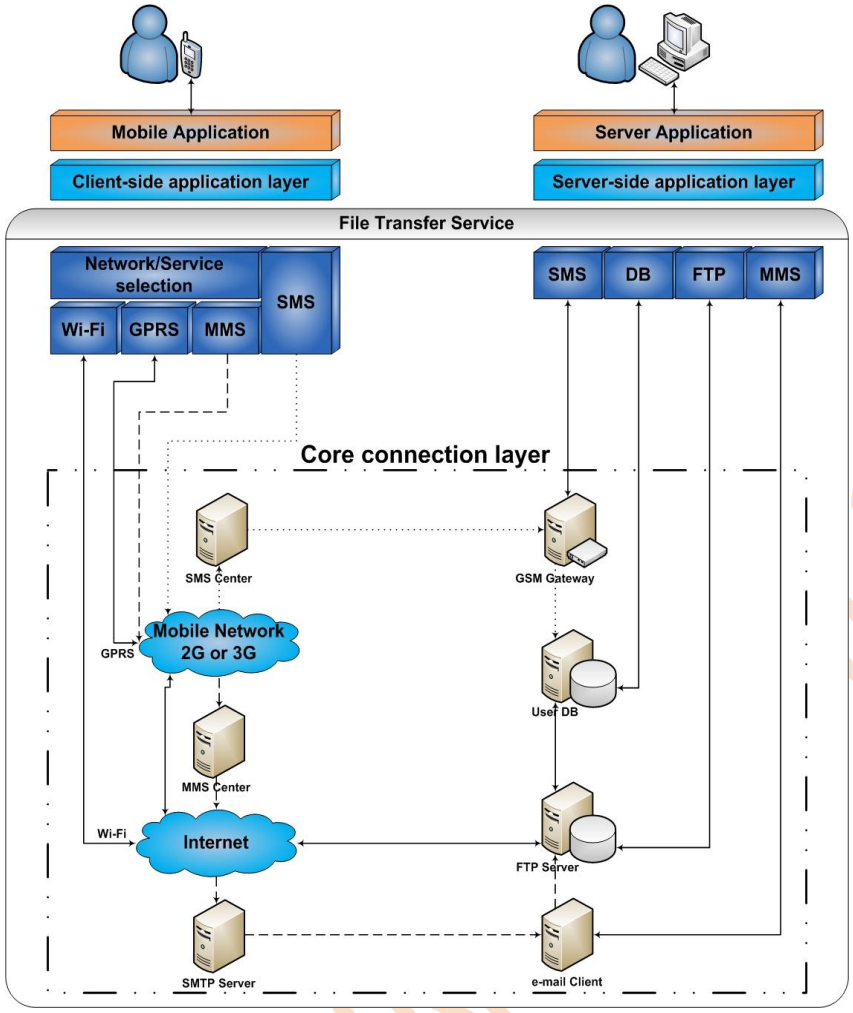


Fig. 1. Layers of the File Transfer Service (FTS) for Mobile Phones with Scare Resources.

C. Server-side application layer (API_ISS)

This layer provides developers an infrastructure (API_ISS) for building a storage server that considers mobile devices. API_ISS includes modules for receiving and sending files through different communication services such as WiFi, GPRS, UMTS and the Multimedia Message Service (MMS).

Different type of servers can be implemented using this API. The particular behavior can be customized, for instance, by developing a web storage service or a file sharing storage server. The Internal Storage Service (ISS) module represents the main module included in the API_ISS. The API ISS is a module that offers a set of methods, which will be used by server-side applications. It contains functions for managing connections and user accounts as well as functions for receiving and transmitting data through different wireless networks. It includes similar wrappers like those located in the client side application layer. It implements the FTP service, which packs all the methods for transferring and receiving files going to or coming from mobile devices. A file storage server developed with the API_ISS could connect with other distributed storage servers building a big virtual disk. This distributed approach allows it to increase its virtual available storage space by integrating the storage space offered by other external storage.

IV. USE CASE

This section describes a use case for the File Transfer Service (FTS). A mobile application named Swapper that uses our FTS was developed. The purpose of Swapper is to send files from a mobile phone to an external storage server when the mobile phone is running out of memory. This process is transparent for users that can always see the complete list of files, even when some of them are not physically kept in the memory of the mobile phone. When a file is required by the user, Swapper requests the missing file from the storage server using any wireless network available or the MMS service, depending on the service priority defined in the configuration file. Automatic swapping allows mobile devices to free storage space in a transparent way. A web storage server was also developed based on FTS. To use our web storage server, users have to be registered before sending files. The registration process can be done through a SMS message or by filling in a web form. Swapper and the FTS framework were developed using the J2ME platform. The configuration, profile and basic libraries needed for our development are: CLDC 1.1 (Connected Limited Device Configuration 1.1.), MIDP2.0 (Mobile Information Device Profile 2.0), WMAPI 2.0 (Wireless Messaging API 2.0), SATSA-Crypto (JSR 177) and PDA Profile (JSR 75).

Due to the fact that Swapper works like a caching system, a file replacement police was included in its implementation. A benchmark was created based on metrics taken from different real mobile phones. The main metrics were: the average storage space in a mobile phone, the average file size, and the average data transfer rate in wireless LAN and WAN networks such as WiFi y GPRS.

TABLE I

COMMON MEASURES IN LIMITED MOBILE PHONES AND PDAS

MEASURE	SIZE
MAX_MEM_SIZE	2GB
MAX_FILE_SIZE	15MB
MIN_FILE_SIZE	469KB
AVG_FILE_SIZE	1.3MB

Table 1 includes measures taken from different mobile phones and PDAs such as: Nokia5530, 7020, N95 and HP iPAQ HW6945, Blackberry 8220.

A collection of more than 1000 files taken from different mobile phones revealed an average file size of 1.3MB. More of them were music, photos and video files. Transmitting files of 1.3MB from different mobile phones to an external storage server connected to Internet using WiFi revealed an average time of 8.3s (about 1.3Mb/s). Using GPRS for the same transmissions, it showed an average time of 148s (about

73Kb/s).

These experiments gave us a realistic average transmission time of wireless networks with different traffic loads. The traffic coming from Internet connections played an important role in these measures. In [13], Enriquez made several tests trying to find an optimal block size for transmitting files between a mobile phone and a data server (PC) using WiFi

and GPRS networks, considering different traffic hours. The optimal block size obtained using WiFi was 4KB (Figure 2) and for GPRS was 96B (Figure 3). However, the GPRS transfer rate showed a high variability in hours with high traffic, resulting in many communications errors. Several block sizes were tested with high traffic to observe the GPRS behavior. GPRS showed less variability when a block size of

64B was used, resulting in a better behavior. This stable behavior motivated the use of block size of 64B in FTS, instead of 96B found in [13].

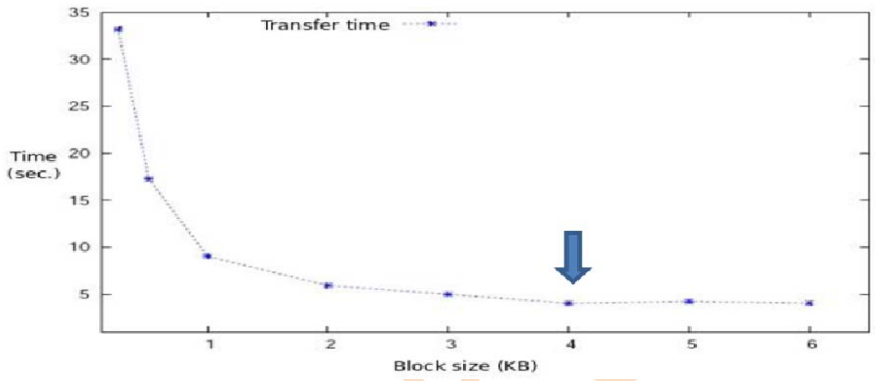


Fig. 2. An optimal transmission block size in a WiFi network

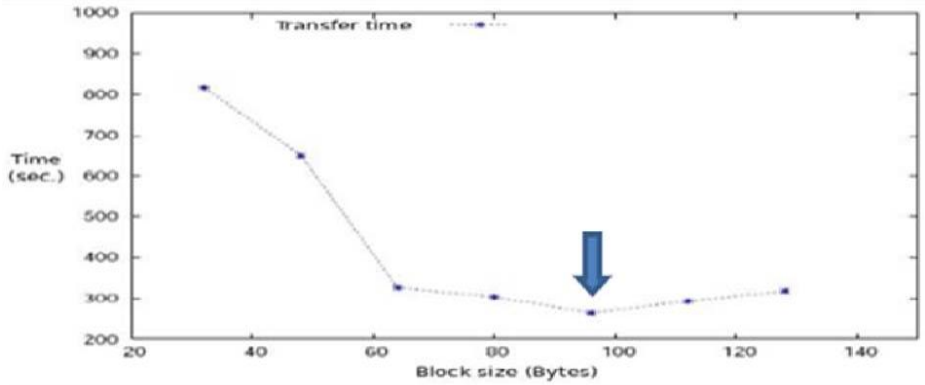


Fig. 3. An optimal transmission block size for a GPRS network

Swapper is a mobile application developed to test our File Transfer Service (FTS). Since Swapper works like a caching service, it had to include a file replacement policy. The main task of a replacement police is to decide which file will be sent to the external storage server to free storage space in the local memory.

In order to find a good replacement policy for Swapper, four algorithms were tested. These algorithms were: LRU (Least Recently Used), LFU (Least Frequently Used), LFS (Largest File Size), and SFS (Smallest File Size). As their names indicate, the LRU policy will send the least recently used file to the external storage server. LFU will send the least frequently used file, LFS the largest file, and SFS the smallest one. If there is more than one file as the largest or the smallest, a LRU police is applied. A benchmark that reproduces a sequence of send/receive operations, using a group of 120 files with sizes ranging from 512KB to 2.5MB was implemented. Thirty one experiments were carried out with different scenarios. Mobile phones were randomly changing between GPRS and WiFi connections, emulating mobile users.

Figure 4 depicts the behavior of the average hit ratio in Swapper using each replacement policy (in 31 experiments). Figure 5 shows the average transmission time

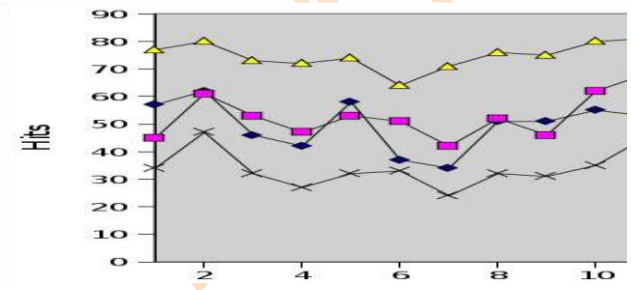


Fig. 4. Average hit ratio obtained by Swapper using different replacement polices

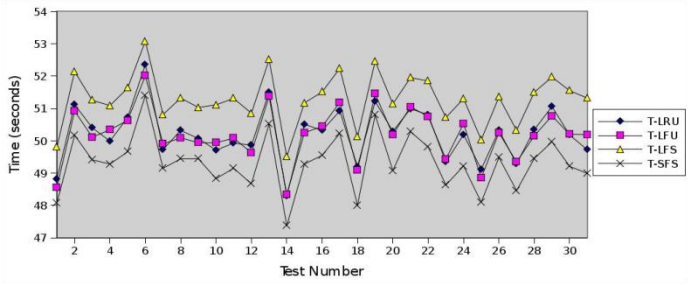


Fig. 5. Average transmission time obtained by Swapper using diferrent replacement policies

As we can see, if Swapper replaces files based on the largest file size (LFS), it obtains the best hit ratio (almost 8 of 10 files were in the mobile phone memory when they were requested by the user). However, it also obtains the greatest average transmission time, which could be a result of transmitting large files using a GPRS connection at the moment of a miss. These results give us some insights to determine if we should prioritize between the quantity of bytes transmitted and the total time using the network, especially in the GPRS/GSM networks. In most cases the wireless communication service is charged based on bandwidth consumption. Figure 6 shows the total bandwidth consumption after running all of the algorithms. These experiments did not show a correlation between the total bandwidth consumption and the average transmission time. Even though the LFU algorithm had low total bandwidth consumption, it did not show low average transmission time. It happened so because most of the time that the mobile application had to send/receive files (using the LRU algorithm) coincided with a GPRS connection.

The results obtained from Swapper gave us information for deciding which algorithm could be implemented as part of our File Transfer Service (FTS), and evaluating if it would be better to include a fixed replacement policy or an adaptive one, taking into account if the cost of the communication service is calculated based on time or bandwidth consumption

In the current implementation of Swapper, users are able to define a cost-based priority list that best fits his or her requirements. This priority list has to be included in a FTS configuration file. Information like depicted in Figures 4, 5 and 6 can help to decide the order of priorities based on the cost of the services in a particular region.

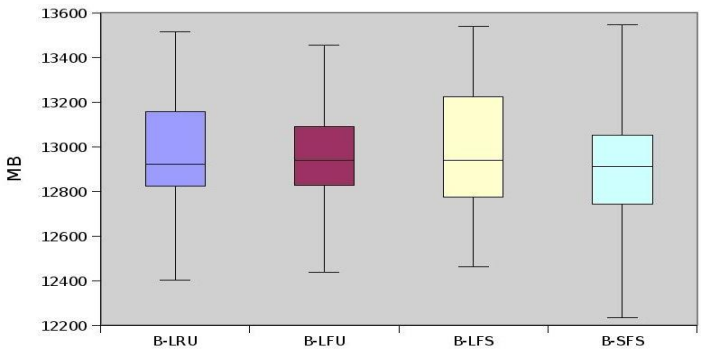


Fig. 6. Total bandwidth consumption obtained by Swapper using 4 different replacement policies.

Another important topic considered in Swapper was information privacy. This topic becomes an important factor when transmitting files from a mobile phone to an external and public storage server. In cases where the storage server does not offer a private place to keep the user information, Swapper can encrypt the files before sending them to the server. In these cases, the external storage server will keep only encrypted files, offering privacy to the information. The encryption process is carried out by using the Java Specification Request 177 (JSR 177, Satsa Crypto). Swapper uses the Advanced Encryption Standard (AES), which is a symmetric key algorithm with a key and block size (it could vary) of 128bits. Experiments for evaluating the impact in the transferring time when using encryption were conducted with Swapper. The objective was to find out the impact in terms of the cost of extra bandwidth and transferring time needed when encryption is applied. For this experiment, the use of the MMS service as another option (included in FTS) to transfer files was tested. To execute this experiment, our test bed had to be redesigned, because the largest public MMS service provider in Mexico has defined a maximum file size for each MMS message of 100KB. Swapper was tested using a group of 100 files with sizes ranging from 80 to 100KB.

TABLE II

AVERAGE FILE TRANSFER TIME WITH AND WITHOUT ENCRYPTION USING FILES WITH AN AVERAGE SIZE OF 90KB

Average File Transfer Time	WiFi	GPRS	MMS
AFTT without Encryption	0.07s	1.23s	32.56s

Table II shows how the encryption process increases the average file transfer time until 3 times in some cases. We can see how WiFi connections are the most affected. This situation is more notorious when small files are transmitted. Since GPRS networks have a low bandwidth and low transmission rates, the resulting impact of the encryption process is lower. The transmission times obtained from the MMS service showed a high variability, because the public MMS server provider in Mexico does not guarantee real time transmissions.

V. ARCHITECTURE

Multi-tier architecture (often referred to as n-tier architecture) is a client–server architecture in which the presentation, the application processing, and the data management are logically separate processes. For example, an application that uses middleware to service data requests between a user and a database employs multi-tier architecture. The most widespread use of multi-tier architecture is the three-tier architecture.

N-tier application architecture provides a model for developers to create a flexible and reusable application. By breaking up an application into tiers, developers only have to modify or add a specific layer, rather than have to rewrite the entire application over. There should be a presentation tier, a business or data access tier, and a data tier.

The concepts of layer and tier are often used interchangeably. However, one fairly common point of view is that there is indeed a difference, and that a layer is a logical structuring mechanism for the elements that make up the software solution, while a tier is a physical structuring mechanism for the system infrastructure.[14] [15]

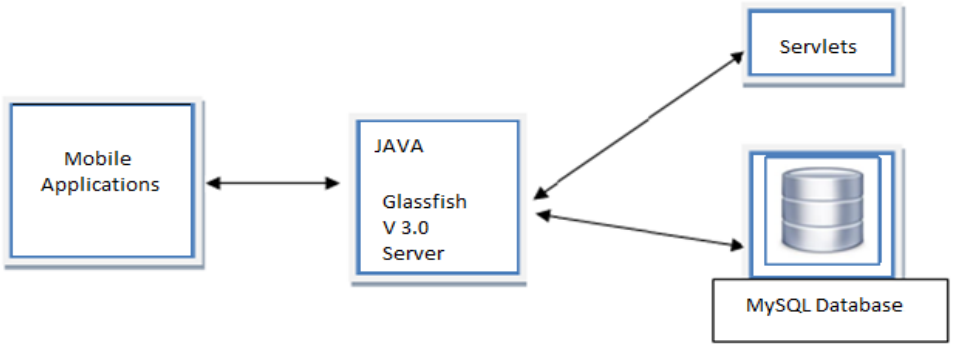


Fig: Main Architecture

This architecture is a 3-Tier Architecture.consisting of Client(Mobile),Server(GlassFishV3.0),BackendDatabase(MySql).So it explain like Requests will be taken from the Client and it will be process in GlassFishV3.0 server the server will contact the database and process the request.The request will be get back to the service method and in turn it will send to the client nothing but mobile. In this architecture we have mobile applications i.e our client side.this client side interacts with server i.e.glassfish server V3.0 and finally this glassfish server stores data in database i.e. MySQL

VI. CONCLUSIONS

As mobile devices evolve, they require more storage capacity to keep the large amount of files that new mobile applications generate.This paper briefly introduced a File Transfer Service (FTS) for mobile applications. The service allows mobile applications to store and request files from an external storage server, taking advantage of available wireless networks and considering issues related to the cost of the service. As a use case, a mobile application

that transparently swaps files between a mobile device and an external storage server was developed, increasing storage space in the former. In this application, different file replacement policies were tested. An important parameter that has to be considered when deciding about a replacement policy is the cost of the service. Results obtained in this work showed that it is not a trivial decision to select a specific replacement policy for a mobile device. Mobile devices do not present a correlation between the bandwidth and time consumption because they do not keep the same wireless connection any time. Service costs are usually defined by wireless WAN providers, and could be based on bandwidth consumption or time consumption, criteria that have to be included in the replacement algorithms. FTS is also prepared for transferring files using the Multimedia Messaging Service (MMS), an additional option to be considered for users when connections are limited. Our experiments revealed that the time for transferring encrypted files could rise the original transferring time until three times. This increase is more evident when small files are transmitted using wireless network with high transmission rates

Acknowledgment

It is indeed a great pleasure and moment of immense satisfaction for us to present a project report on 'An External Storage Support for Mobile Applications with Scarce Resources'.Amongst a wide panorama of people who provided us inspiring guidance and encouragement ,we take the opportunity to thank those who gave us their indebted assistance. We wish to extend our cordial gratitude with profound thanks to our Guide,Prof.S.M.Bhadkumbhe for her everlasting guidance.It was her inspiration and encouragement which help us in completing of our project.And our sincere thanks to all those individuals involved both directly and indirectly for their help in all aspects of the project.

REFERENCES

[1] Satyanarayanan M., Kistler J.J., Kumar P., Okasaki M.E., Siegel E.H., and Steere D.C.. Coda: a highly available file system for a distributed workstation environment. IEEE Transactions on Computers 39(4):447-459, April 1990.

[2] Kistler J.J. and M. Satyanarayanan. Disconnected operation in the Coda file system. ACM Transactions on Computer Systems lO(1): 3-25, February 1992.

[3] Satyanarayanan m.. Mobile Information Access. IEEE Personal Communications, 3(1):26–33, Feb. 1996.

[4] Demers, A., Petersen, K., Spreitzer, M., Terry, D., Theimer, M., Welch, B.: The bayou architecture: Support for data sharing among mobile users (1994).

[5] Mascolo, C., Capra, L., Zachariadis, S., Emmerich, W.: Xmiddle: A data-sharing middleware for mobile computing. Int. Journal on Personal and Wireless Communications 21(1) (April 2002) 77-103.

[6] Lui, J.C.S., So, O.K.Y., Tam, T.S.: NFS/M: An open platform mobile file system. In: ICDCS '98: Proceedings of the The 18th InternationalConference on Distributed Computing Systems, Washington, DC, USA, IEEE Computer Society (1998) 488.

[7] Atkin, B., Birman, K.P.: Network-aware adaptation techniques formobile file systems. In: NCA '06: Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications, Washington, DC, USA, IEEE Computer Society (2006) 181-188.

[8] GSpaceMobile:(Mar.2009).Available at:https://www.ibomobi.com/home/.

[9] Emoze™: (Aug. 2009) . Available at: http://www.emoze.com/.

[10] Boulkenafed, M., Issarny, V.: A middleware service for mobile ad hoc data sharing, enhancing data availability. In: Middleware '03: Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware, New York, NY, USA, Springer-Verlag New York, Inc. (2003) 493-511.

[11] Belimpasakis, P., Luoma, J.P., Börzsei, M.: Content sharing middleware for mobile devices. In: MOBILWARE '08: Proceedings of the 1st international conference on MOBILe Wireless MiddleWARE, Operating

Systems, and Applications, ICST, Brussels, Belgium, Belgium, ICST

(Institute for Computer Sciences, Social-Informatics and

Telecommunications Engineering) (2007) 1-8.

[12] Bodic, G.L.: MOBILE MESSAGING TECHNOLOGIES AND SERVICES SMS, EMS and MMS. Second edn. (2005).

[13] Enríquez, J.A.V.: A vertical handover platform for applications on mobile devices. Master's thesis, CINVESTAV (2009). Available at:

http://www.tamps.cinvestav.mx/sites/default/_les/tesis 1.zip.

[14] Deployment Patterns (Microsoft Enterprise Architecture, Patterns, and Practices)

[15**]** Fowler, Martin "Patterns of Enterprise Application Architecture" (2002). Addison Wesley.

IT 027

Best Implementation and High Performance Storage Virtualization Architecture

Sachin V. Joshi

Department of MCA

P.R.M.I.T.& R. Bandera, Amravati

joshi7.sachin@gmail.com

Miss. Shrutika A Hajare

Department of MCA

P.R.M.I.T.& R. Bandera, Amravati

shrutikahajare@yahoo.com

Devendra R. Bandbuche

Department of MCA

P.R.M.I.T.& R. Bandera, Amravati

devendrabandbuche@gmail.com

Abstract:

This paper presents a best implementation and

efficient performance for storage methods and a virtualization method that provides the storage management protocols or tool. This technique aims improving storage problems in virtualization architecture. Currently, storage virtualization is implemented at three architectural levels: (a) at the storage sub-system, (b) in the host and (c) in the storage Network-based virtualization is being implemented in two major architecture that are Symmetric Virtualization and Asymmetric

Virtualization. It provides maximum number of devices available on the virtual architecture, to have limitations in bandwidth, it creates central location for difference devices. The storage management benefits offered by existing techniques without the corresponding drawbacks.

1. INTRODUCTION

The purpose of storage virtualization is to separate physical line issue of cost. The implementing way is that an abstract layer is added between physical storage equipment and the customer side that can provide a simple and universal storage pool to customer through uniting various storage employments. The network- layer storage virtualization in SAN (storage area network) is implemented by a special device. This special device is a server platform that contains storage virtualization management and application software. We call the server that manages storage network as Manager (storage virtualization controller). The abstraction, or virtualization, enables many new benefits and capabilities that have been highlighted as part of the benefits with Storage Area Networks. Enabling the ability to deal with storage in abstraction where the details of configuration, device specific features and characteristics are transparent to the server software allows use of differing types of storage, commonality of management tools, and exploitation of a common set of features.

2. CHALLENGES FOR IT TODAY WITH DISK STORAGE

The growth of data storage has to be the most obvious and pressing problem facing IT organizations. The growth being experienced is unprecedented and has led to a multitude of specific issues. Obvious problems are the costs associated with the physical purchase of more storage, the physical space and environmental costs, and the operational time to bring storage into service. The longer-term challenge is the added administration cost for that storage with continued growth, this becomes a compounding problem. This growth of storage exacerbates another severe problem: the availability of people with the skills to manage storage. With a given set of tools, a storage administrator can manage a specific amount of storage. This means that with storage growth, more administrators will be required. This cost is quantifiable but may be moot if there are not enough skilled people to get the job done. The increased competitiveness and the amount of information have led to very complex operational environments. There is more demand for “data integration” where information from multiple sources and systems, in differing formats, can be correlated and presented in a

unified, informative manner. This expectation is today and is immediate. Achieving it requires significant planning and work. All these challenges result in a bottom line issue of cost. They represent significant costs in the administration of storage systems and in the infrastructure implementation to be able to meet increased demands.

3. SOLUTION FOR THESE CHALLENGES

Accommodating the growth storage can be non-disruptively added and data can be migrated as necessary without operational impact. This means that the storage infrastructure has to be capable of allowing that and the operating system and applications on servers can transparently handle the addition of storage. Storage Area Networks promise the infrastructure for this but it is the virtualization or abstraction of storage that allows it to be transparent and non-disruptive.

Virtualization allows policy-based management of storage to be implemented in a practical manner and allows the management tools to be consistent across varying devices. This means that storage administrators would have to learn only one set of tools to manage storage and that purchasing decisions would be based on costs, reliability, etc. at the planned time of need rather than a decision for a predetermined vendor set done at an earlier time and driven by operational impact required to change. In that very complex environment, storage administrators can typically administer greater amounts than in an environment with multiple servers, operating systems and device types. Being able to administer more storage and to be more quickly trained on a specific, common set of into management tools helps address the availability problem of administration people skilled in storage administration.

4. THE BEST IMPLEMENTATIONS OF

VIRTUALIZATION

There are several choices for implementing virtualization and vendors have implemented a variety of solutions. Understanding the possible implementations and what is involved in doing and using the implementation should lead to a clear choice of where virtualization should be done. In fact, most systems in

production today are utilizing some form of block-level virtualization in which a logical block address from the host is mapped in some way to physical storage assets. The major change is that block-level virtualization is increasingly seen as a key technology for the implementation of new services and the reimplementation of some older, but previously proprietary, services in the storage network.

4.1 VIRTUALIZATION AT THE HOST LEVEL

One method of virtualization is via storage management software that runs at the server level. The main advantage of this method is that it enables multiple storage subsystems to work in parallel with multiple servers. A key difficulty with this method is that it assumes a prior partitioning of the entire SAN resources (disks or LUNs) to the various servers. Virtualization is only performed on pre-assigned storage, losing a key advantage of SANs as well as the independence of volumes from servers. Typically, entire LUNs are assigned to specific servers, thus limiting the amount of servers that can use the same storage subsystem. Further, a virtual volume created over the storage space of, say, two LUNs will not be

easily moved to another server, especially if there are other

volumes created over the same LUNs. Virtualization at the host level typically also require augmenting the management function with some parallel mechanism of zoning and LUN masking and relying on LAN connectivity for synchronization between servers, which may affect the reliability of the entire SAN.

4.2 VIRTUALIZATION AT THE STORAGE SUBSYSTEM LEVEL

Even though you may not have known it, your storage arrays may have been performing storage virtualization for years. Features including RAID, snapshots, LUN (a logical unit number) masking, and mapping are all examples of block-level storage virtualization. Storage-based virtualization techniques are equally applicable in SAN and DAFS (direct access files system) environments Storage-based virtualization is typically not dependent on a specific type of host, allowing the array to support heterogeneous hosts without worrying about variance of host operating systems or applications. Also, storage-based RAID systems deliver optimum performance in relation to their hardware because features like caching can be uned to the specific hardware. The downside to this approach is that the

storage virtualization functions are typically confined to a single array; for example, the source volume used for a snapshot and the snapshot itself are maintained on the same array, making the snapshot useless in case of hardware failure. In some cases, virtualization functions extend across multiple arrays, or a cluster of arrays or controllers; however, these solutions are typically restricted to a single-vendor implementation.

Frequently, host- and storage-based virtualization is

combined, adding the flexibility of the host-based LVMs to the performance of hardware-assisted RAID. For example, the host-based LVM (logical virtual masking) can use multiple RAID-5 LUNs (a logical unit numbers) to create virtual volumes spanning multiple disk arrays. In addition to simply striping across LUNs from multiple arrays, LVMs in the host can also mirror (with striped mirrors or mirrored stripes) volumes across multiple arrays. Host-based LVMs are also used to provide alternate path fail-over in most environments because the host is the only device in the I/O chain that knows for certain whether its I/O completed. Similarly, LVMs may also implement load balancing on the access paths to the storage to increase performance.

4.3 NETWORK –BASED VIRTUALIZATION

This is the most interesting approach and the most likely to win out due to its neutrality towards both storage and servers. A key requirement from Storage Virtualization is “making disparate storage look and behave like a common storage resource”. Network base virtualization assures vendor neutrality. Network-based virtualization is being implemented

in two major architectures:

1. Symmetric Approach (In-Band irtualization)

2. Asymmetric Approach (Out-of-Band Virtualization)

4.3.1 SYMMETRIC APPROACH (IN-BAND

VIRTUALIZATION)

With an in-band appliance(s) located in the data path between hosts and storage, all control information (metadata) and data pass through it. Another term used to describe this process is store and forward. To the host, the appliance looks and behaves like a storage array (e.g., an I/O target) that presents logical volumes. To

the storage, the appliance looks and behaves like a host, issuing read and write requests (e.g., an I/O initiator) that are indistinguishable from I/O request generated by conventionally attached hosts. The general architecture of symmetric approach (sometimes referred to as In-the-Data-Path) This approach calls for the appliance (a computing platform and associated memory) to be installed between the storage users (i.e. the servers on the SAN) and the storage resources, at least in terms of data flow. The key drawback of the symmetric approach to virtualization is that it creates a SAN bottleneck and thus limits the SAN performance and scalability and significantly complicates the design of large scale highly available configurations. The symmetric virtualization concept requires that all data from all application servers pass through one single computer. To avoid serious performance difficulties, this computing platform needs to be capable of sustaining the throughput of the entire SAN, typically resulting in expensive hardware configuration. Even with high performance hardware, scalability remains an issue as the symmetric appliance has a fixed maximum bandwidth.

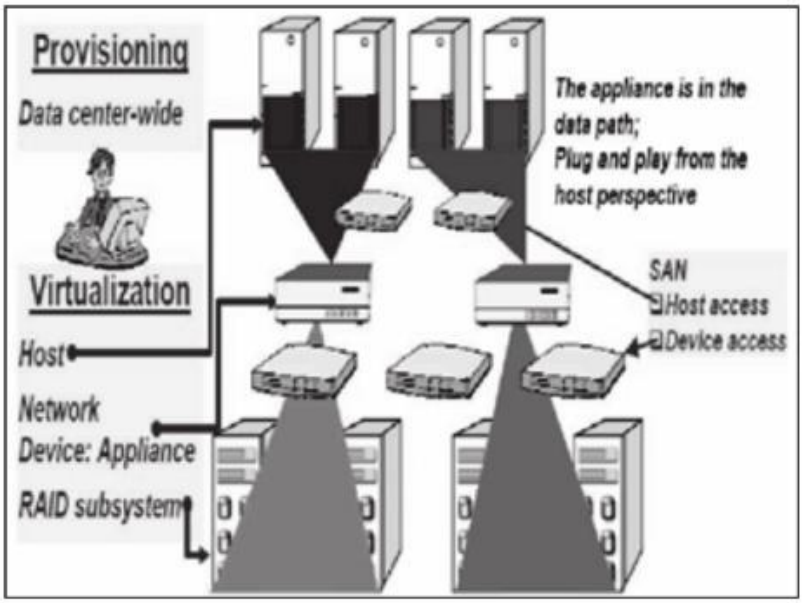
As an example, in a SAN with 50 application servers, the throughput required for each can reach 100 MB/sec. This value * 50 servers = 5 G bytes/sec. While it is fairly easy to install multiple RAID subsystems with aggregated performance exceeding this requirement, a typical symmetric appliance based on a standard Pentium III processor with a 64/66 PCI bus can deliver no more than 520 MB/sec (that is 260 MB/sec sustained rate as data needs to enter and exit through the PCI (Peripheral component internet) bus on the appliance – a far cry from the specified requirement. Needless to say, as the SAN becomes larger, this problem becomes more acute.

As an example, in a SAN with 50 application servers, the throughput required for each can reach 100 MB/sec. This value * 50 servers = 5 G bytes/sec. While it is fairly easy to install multiple RAID subsystems with aggregated performance exceeding this requirement, a typical symmetric appliance based on a standard Pentium III processor with a 64/66 PCI bus can deliver no more than 520 MB/sec (that is 260 MB/sec sustained rate as data needs to enter and exit through the PCI (Peripheral component internet) bus on the appliance – a far cry from the specified requirement. Needless to say, as the SAN becomes larger, this problem becomes more acute.

The hardware cost to construct a symmetric appliance that will support 20 servers with reasonable performance could reach $25,000 and more; this cost is at least doubled if a High Availability configuration is required.

Some symmetric designs are attempting to circumvent the performance problem through adding a cache facility to the appliance. This approach further loads the appliance's memory subsystem. For many reasons cache is much more effective when it is distributed (i.e. part of the RAID subsystem) rather than centralized (i.e. part of the storage management appliance).

Figure 1 – In-Band Virtualization



4.3.2 ASYMMETRIC APPROACH (OUT-OF-BAND

VIRTUALIZATION)

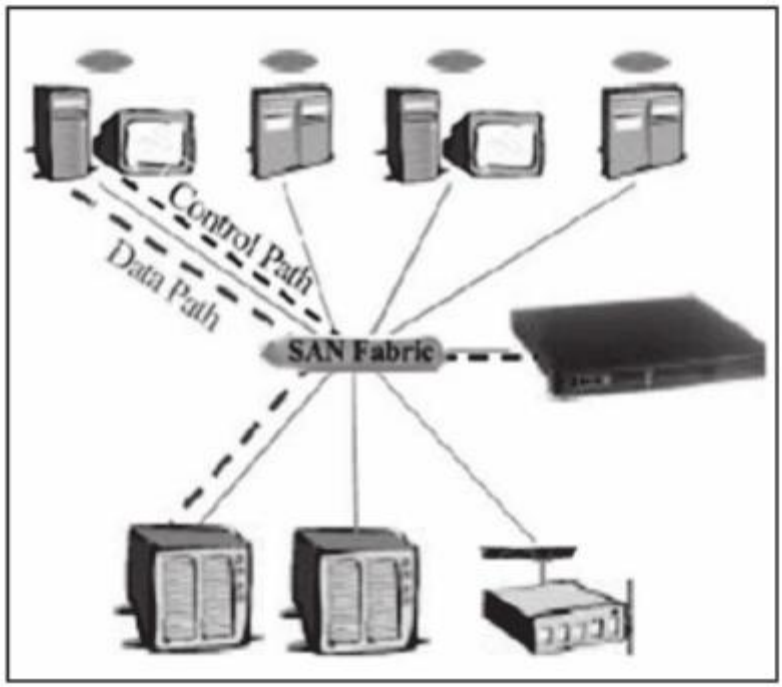
Asymmetric approach network-based virtualization, this approach is now available on the market. The OOB (out of band management functions) appliance communicates with the host systems via Ethernet or Fiber Channel in order to resolve requests for LUN access, volume configuration, etc. Because hosts must contact the appliance in order to gain access to storage, they must be aware that storage is being virtualized, unlike in-band approaches in which the virtualization function is transparent to the hosts. As a result, OOB virtualization architectures require a virtualization client or agent in the form of software or special HBA (host base adapters) drivers installed on each host. This agent receives information on the structure and properties of the logical volumes as well as the corresponding logical/physical block mapping information from the appliance.

This method uses a combination of an Appliance and Agents to create and manage

virtual volumes while enabling direct data transfer between server and storage subsystems (see Fig. 2). By having multiple storage subsystems working in parallel with multiple servers, total performance is increased up to the maximum of the FC fabric bandwidth or even more. In the Asymmetric approach (sometimes referred to as Out- of-The-Data-Path), the handling of metadata is done separately from the data path. The appliance serves as a Metadata Center that "sees" the physical storage and allocates virtual volumes, while an Agent at each of the servers on the SAN is responsible for the actual virtual volume mapping. The Agent retrieves the volume configuration from the appliance and presents virtual volumes to the operating system as if they were disk drives. When the operating system sends an I/O to the virtual volume, the Agent intercepts the I/O, translates the volume's logical address to the physical address, and sends the I/Os directly to the storage devices.

Figure 2 – Out-of-Band Network Based Virtualization

This architecture maintains the flexibility of symmetric virtualization without incurring the degradation of performance or the high cost of hardware. The appliance can be a small and inexpensive unit, because it does not have to handle actual data transfer.



5. CHOOSING WISELY

It should be obvious that a virtualization solution that spans across systems in the enterprise provides the greater gains in benefits provided by virtualization. Individual solutions will not provide the same business value to customers that an enterprise-wide solution is capable of doing. The solution needs to be comprehensive in providing abstracted storage for all servers connected in the Storage Area Network. The business value seen in reduced administrative costs and the realization of the benefits of virtualization should be able to be demonstrated to reaffirm the choice for the virtualization solution. A comprehensive solution that virtualizes disk storage across the enterprise rather than at a specific subsystem has the potential for the greatest gains in reduction in administrative costs and enabling of enterprise-wide features. The real measure is the value brought to the business for the particular solution. A cost-benefit study showing the improvements in administration of storage as part of total cost of ownership should be illuminating for making a decision.

5.1 ABBREVIATIONS

Storage Area Network (SAN), Fiber Channel (FC), Linea Tape Open (LTO), Non-volatile random access memory (NVRAM), Logical Unit Number (LUN).

Table: Analysis of Various Virtualization Methods

Method	Host Based	Storage Based	Symmetric Virtualization (In-Band Virtualization)	Asymmetric Virtualization (Out-of-Band Virtualization
Universal	-	-	+	+
Single point of management	-	+	+	+
Flexibility	-	-	+	+
Performance	+ +		-	+
SAN scalability	Poor	Poor	Poor	High
Relative management cost	Medium	High	High	Low
Security	-	+	+ High security level since Server cannot "see" the storage	SVM Increases Security level by Smart Algorithms of "Automatic Switch Zoning".

6. CONCLUSIONS

Virtualization of disk storage will provide such compelling benefits that it will inevitably be a common implementation for computer systems. The advantages for enterprise-wide virtualization will lead to that implementation becoming dominant over time. The best way to implement storage virtualization is by the asymmetric architecture using a

combination of SAN appliance and agents. This method provides significant scalability, performance, reliability and price advantages over alternative architectures.

Analysis of various virtualization methods better storage networking management for next generation performance management protocols most critical elements for business applications it its availability to its own storage data.

Correlate business application performance information across servers and storage networks. Providing metrics and monitoring for proactive trending of performance.

REFERENCES

1] Robert Spalding, Storage Network :The Complete Reference. Tata McGraw-Hill 2003.

2] Sean Campbell and Michael Jeronimo, "An Introduction to Virtualization", 2006 Intel Corporation.

3] A New Approach to Storage Virtualization, White Paper, Spinnaker Networks.

4] The Value of Storage Virtualization, White Paper, Falcon Store Software, 2001.

5] Virtualization of Disk Storage, WP0007-2, The Evaluator Series, Evaluator Group, Inc., February 2002.

6] P.M. Chen, E.K. Lee, G.A. Gibson, R.H. Katz and D.A. Patterson, RAID: High Performance Secondary Storage, ACM 7] Computing Surveys, Vol. 26, No. 2, June 1994, pp.145-185 G. Schultz, SAN and NAS: Complementary Technologies, MTI Enterprise Storage, 2000.

8] M. Krueger, R. Haagens, et al., IETF RFC 3347, Small Computer System Interface Protocol over the Internet (ISCSI) Requirements and Design Considerations.

9] ANXI X3.297- 1997, Fibre Channel Physical and signaling Interface-2.

10] NCITS T11/Project 1508-D/Rev 6.01, Fiber Channel Switch Fabric -3 (FC-SW-3).

11] Gary Field, Peter Ridge, et al. "The books of SCSI:I/O for the new millennium", 2nd edition, 1999, No Starch Press.

12] Tom Clark, "IP SANs – A Guide to iSCSI, iFCP, and FCIP Protocols for Storage Area Networks", 2002, Pearson Education, Inc.

13] Julian Satran, Kalman Meth, et al, Iscsi IETF Internet Draft: ISCSI, draft-ietf-ips-iscsi-20.txt

14] M. Rajagopal, et al., IETF draft: Fibre channel over TCP/IP (FCIP), draft-ietfipsfcovertcpip12.txt.

15] DavidAPatterson,GarthGibson, and Randy H Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)", Proceeding of the 1988

ACM SIGMOD international conference on Management of data. 1998, Chicago, Illinois, United States.

IT 028

Recognition of Guilty Agent in Information leakage

Poonam Sahoo

University of Pune, India
e-mail: poonam.sahoo00@gmail.com

Kanchan Garud

University of Pune, India
e-mail: garudkanchan11@gmail.com

Sarika Kate

University of Pune, India
e-mail: katesarika2@gmail.com

Sharayu Padmane

University of Pune, India
e-mail: sharayupadmane@gmail.com

Abstract— While doing business, often sensitive information is required to be shared by supposedly trusted agents. The owner of the data is called distributor founds the data in some unauthorized place and now the distributor must access the likelihood that the data was leaked by one of the agent as opposed to have been gathered by independent means. We have data allocation strategies that improve the probability of detecting leakage. In this method, we inject some realistic but "fake" records which are used in calculating the probability of Guilty agent.

Keywords- Fake records, Guilty agent, Leaked dataset, sensitive data, Probability of guilty agent ,Allocation strategies, data privacy, leakage model.

Introduction

While doing business, crucial and private data is sometimes handed over to the supposedly trusted third parties. For example, a Insurance Company may

give its customer data for further processing to other agents like what is the average salary of people who are getting themselves insured for amount greater than Rs.100000. We call the owner of the data *distributor* and the supposedly trusted third parties *agents*. These agents may behave in such a way so as to achieve their own objectives even at the cost of others.

If any agent gives this information to other agents without informing the distributor then the data is said to be leaked by the agent. Our goal is to detect leakage by such agents and also to identify the guilty agent. For this purpose we have data allocation strategies that use some realistic but fake records. These fake records help in the calculation of probability for each agent being guilty.

Drawbacks of the current methodology

Traditionally, leakage detection is handled by watermarking which is method of embedding a unique code on each of the distributed copy. When this copy is later discovered in the hands of an unauthorized party, the leaker can be identified.

However, there are two major disadvantages of the above algorithm:

1. It involves some modification of data i.e. making the data less sensitive by altering attributes of the data. This alteration of data is called perturbation. However in some case it is important not to alter the original distributed data. For example, if an outsourcer is doing the payroll, he must have the exact salary. We cannot modify the salary in this case.

2. The second problem is that these watermarks can be sometimes destroyed if the recipient is malicious.

Objective of Proposed system

The main objective of the paper is to analyze the guilt of every agent which can be responsible the leakage.

The paper mainly deals with customer data which is important for companies like Life insurance companies. If data leakage happens, they can't afford to lose their valuable customers through which they earn their income as well as reputation.

The whole system is divided into three parts:

1. Data allocation strategies

In our system, we have such data allocation strategies [3] that involve addition of some realistic but fake objects while distributing data to the agents in order to improve the probability of finding the guilty agent.

2. Mail Detection system

Our project basically deals with customer information which is crucial for a company like life insurance agency. When such data is leaked, the third party tries to establish contact with these leaked customers through mail. Our mail detection system keeps a track of these fake email-ids and when an unidentified mail having advertisement content arrives, the system informs the administrator about it.

3. Probability calculation

In the final stage, when the threshold value of unidentified mails is crossed, the system calculates the probability of each agent being guilty with the help of these fake records [2] and the one having the highest probability is the Guilty agent. The threshold value is the minimum value required to trigger the system for the calculation.

Problem defination

The distributor has to distribute data in an efficient manner such that leakage can be detected.

There are many cases where alteration to the original data can't be done. In such cases, we can add some realistic records similar to the dataset which don't exist in reality.

For example, In a Life Insurance Agency the distributor can't alter the personal and contact details of the customer. In order to watermark such data, some fake customer records are acceptable since no real customer matches with this record.

There are two major constraints in this system:

1. The distributor has to satisfy agent's request by providing them with the number of objects they request. The distributor may not deny serving an agent request and may not provide different perturbed versions of the same objects.

2. The second constraint is to allocate data in such a way that the guilty agent is tractable.

Problem Setup and Notation

1) Entities and Agents A distributor owns a set

T = ft1; : : : ; of valuable data objects. The distributor

wants to share some of the objects with a set of agents

U1;U2; :::;Un, but does not wish the objects be leaked to

other third parties. The objects in T could be of any type and size, e.g., they could be tuples in a relation, or relations in a database. An agent Ui receives a subset of objects Ri _ T, determined either by a sample request or an explicit request:

- Sample request Ri = SAMPLE (T;mi): Any subset of records from T can be given to Ui.
- Explicit request Ri = EXPLICIT (T; condi): Agent Ui receives all the T objects that satisfy condition.
- MODULE DESCRIPTION
- Database Maintenance
- The sensitive data which is to be distributed to the agents is maintained in this database.

- Agent Maintenance
  The registration details provided by the agent as well as the data which is given to it by the distributor is maintained.
- Addition of fake Objects
- The distributed is able to add fake objects in order to improve the effectiveness in detecting the guilty agent.

- Mail Detection system
- The monitoring of mails on the fake records distributed is done by this system.

- Data Allocation
- In this module, the original records fetched according to the agent's request are combined with the fake records generated by the administrator.

- Calculation Of Probability

- In this module, the each agent's request is evaluated and probability of each agent being guilty is calculated.

VI. SYSTEM FLOW

In the beginning, agent registers into the system by entering its personal details and the request of data. The system then extracts the requested data from the main database and performs the addition of fake records according to the request. It then provides the data to the agent.

If any of the agent leaks the data to some unauthorized vendor, then the vendor will try to establish a contact with the customers by sending them advertising mails. The job of the mail detection system is to monitor these incoming mails on the email addresses of the fake records continuously.

If the system detects unauthorized mail crossing the threshold value, then it starts its process of probability calculation. The *threshold value* is the minimum number of mails which have to be detected to trigger the calculation. It will check the presence of these fake records in each of the agent and accordingly will evaluate the probability of each agent being guilty. The agent having the maximum probability will be the guilty agent.

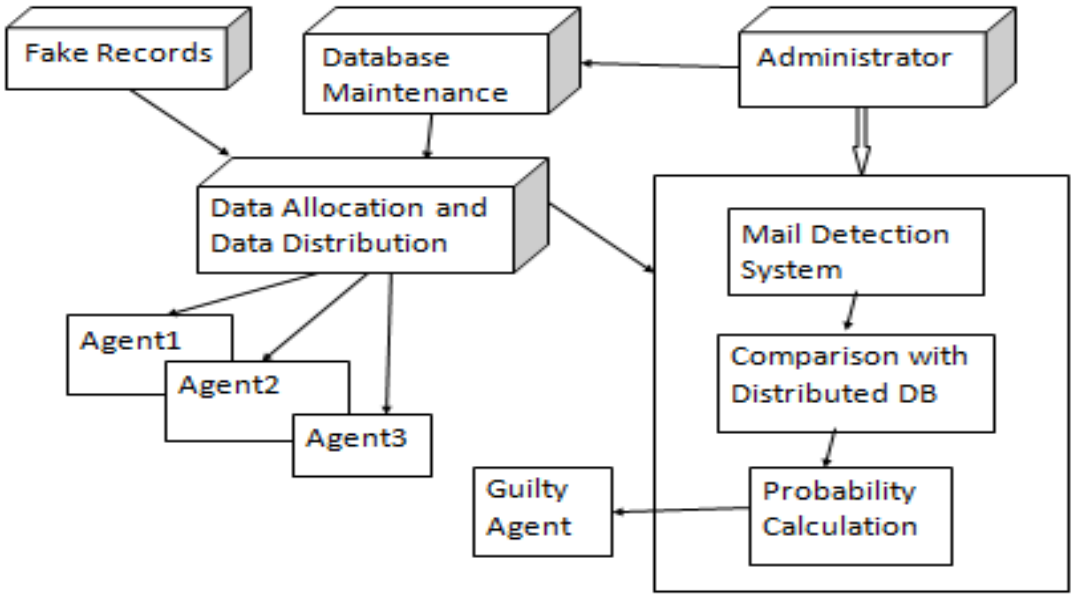


Fig 1. Architecture view of the system

VII. CONCLUSIONS AND FUTURE WORK

Scope of the above system can be extended by handling the generation of fake records dynamically according to the agent's request. Accuracy of the calculation of the probability becomes more, if we have a method to get the complete dataset.

In real world, there is always a need to handover crucial data to third parties and we can't distribute such data by watermarking each and every object which also involves modification. Hence, we have to implement a system where data is not modified yet watermarked which will signify the ownership of the distributor.

References

[1] Data Leakage Detection, Panagiotis Papadimitriou, Hector Garcia -Molina (2010), IEEE Transactions on Knowledge and Data Engineering, Vol 22,No 3

[2] R. Agrawal and J. Kiernan, "Watermarking Relational Databases "Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), VLDB Endowment, pp. 155-166, 2002

[3] P. Buneman and W.-C. Tan, "Provenance in Databases," Proc.

[4] Y. Cui and J. Widom, "Lineage Tracing for General Data Warehouse Transformations," The VLDB J., vol. 12, pp. 41-58,2003.

IT 029

Instant Academic Notices on Phones

Satish Kamble
Computer Engineering Department
Pune Vidyarthi Griha's College of Engineering.and Technology, University of Pune India

Prachi Tongaonkar
Computer Engineering Department
Pune Vidyarthi Griha's College of Engineering.and Technology, University of Pune India
prachitongaonkar@gmail.com

Remya Mogayan
Computer Engineering Department
Pune Vidyarthi Griha's College of Engineering.and Technology, University of Pune India
ramyaa2112@gmail.com

Shreya Kadam
Computer Engineering Department
Pune Vidyarthi Griha's College of Engineering.and Technology, University of Pune India
shreyakdm@gmail.com

ABSTARCT

In this paper we propose a system that automates the traditional notification system. This system aim towards providing a centralized architecture that is beneficial for an academic institution by using mobile and web technology. It integrates a server with a central database to be accessed by the mobile applications, programmed for Google Android platform. The architecture of the system is client server. The administrator will update information on web server. The client who intends to use an Android smart phone will be notified on his mobile application by a prompt. He can further request for data download. The system will deal with information in academic institution like time tables, events, schedules, news, prompt notices, etc. The system when employed in an

academic institution can be useful for prompt messaging and circulation and distribution of information.

Keywords— android, database server, download, push technology, smart phone, upload.

I. INTRODUCTION

The traditional approach of notification system in colleges and schools experiences some drawbacks. The drawbacks of existing system are vast like lack of circulation of information, prompt messaging system and revised updates. It is also time critical system and creates clutter and confusion. The system also not very green (use of papers).

The existing system is time critical, improper circulation of information, has physical barriers. So an alternative is to automate the system with the use of modern technology. The use of mobile phone and Internet can provide the solution to the problem.

Firstly, a web portal is created for an academic institute. On this web portal notices can be uploaded. It will be sent on mobile phones. After the completion of the web portal, the mobile app is incorporated to receive the sent notices on phones.

The system that is proposed will eradicate the drawbacks of the existing system. The system to be implemented is the system that is automated and fast and reliable.

It is been observed that use of phone has increased substantially. The Internet facility is also easily available in the academic institute. The students as well as teachers use the phones in large frequency. So idea was to use this demanding factor in replace the existing time consuming system with more reliable and fast service.

II. OVERALL IDEA

The product proposes an approach towards overcoming these drawbacks. Since the product uses the phones which is easily available or possessed by everybody. An authorized or privileged user will just upload the notice by login into portal. He will select the intended group of users of notice. The notice will be the pushed to latter. Thus the proposed system is less time and effort consuming.

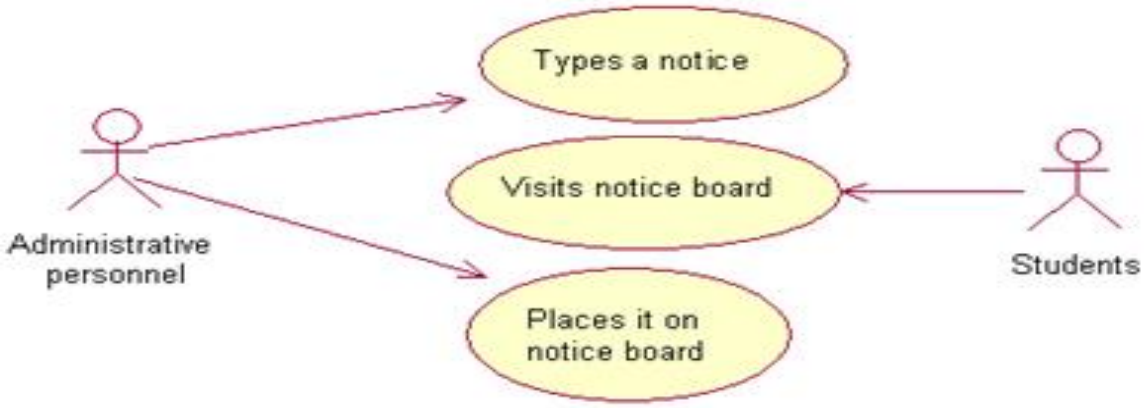


Fig 1: Traditional approach of Notice boards

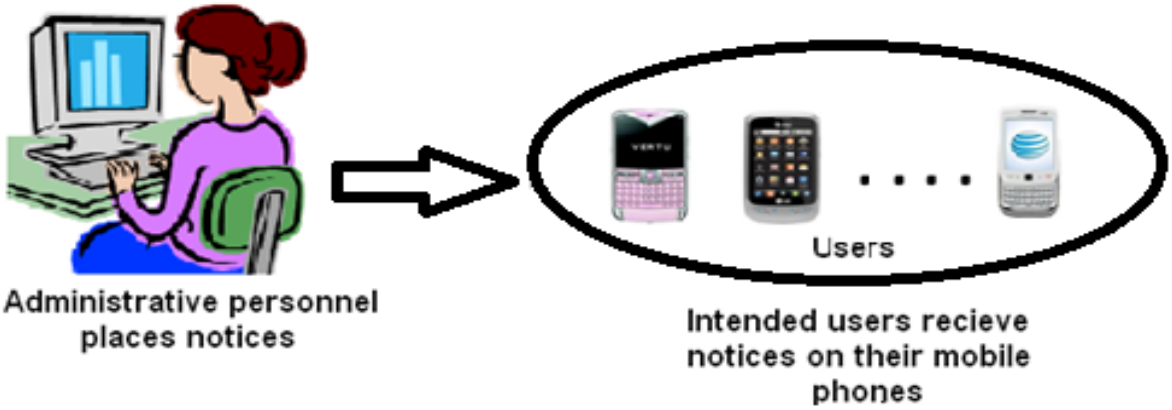


Fig 2 : Proposed approach of notice boards

A. **Product Plan**

The system to be developed is intended to have some characteristics. We plan that a privileged user can send notices to all users or a group of users. There will be provision to create groups can be made according to their stream, course, class or designation. Various privileges are provided to authorized users. All types of documents like time-tables, schedules, assignment lists, etc that are related to any academic institution. The task of uploading notice may take few minutes. The circulation of correct notice is done to all users. Due to prompt messaging system, it becomes time saving system. It eradicates the problems occurred due to physical barrier. A mobile app will really make it simpler. It does not involve much maintenance cost

B. User Classes and Characteristics

All the end users of the system are the people related to an academic institute like students, lecturers, administrative personnel, etc. Considering their roles and authorities the users are classified into following categories:

a) End User

This category of users particularly includes the students studying in that academic institute. This is main category of users to whom actually the notices are intended for. According to the course, the students are appearing for, they can be classified further. All the notices regarding time tables, schedules, exam dates, any modification in any lecture timing, etc are important for this category of users.

b) Intermediate User

This category of user includes the lecturers, head of the departments, principal, administrative personnel, training and placement officials, librarians, etc. This category of user has privilege to post a notice and also they are subject to receive any relevant notice. They are provided with some privileges and authorities according to their designation.

c) Administrator

This is the user who has the maximum privilege. He is responsible to add or remove any user, give privileges to other user, create and manage groups, manage databases, assign temporary privileges to acting users, etc. He has control on overall system. He has to also manage the of how much duration will the notice and its related content will be available.

III. SYSTEM OVERVIEW

A. *Functionalities*

- We need to develop the a efficient system that will prove better than the existing system. For that we need to consider some points:
- *Creation of all types of user accounts and proper validations on them.*
- *Providing correct privileges to all users.*
- *Administrator should be able to create various groups.*
- *Administrator can manage groups by adding, deleting members and modifying their privileges.*
- *Creation and management of database by including their account details and privileges.*
- *All the notices are stored along with their details like notice id, reference number, subject, etc.*
- *Synchronization between Server and mobile app.*
- *Deployment of Android app.*

B. System's Architecture

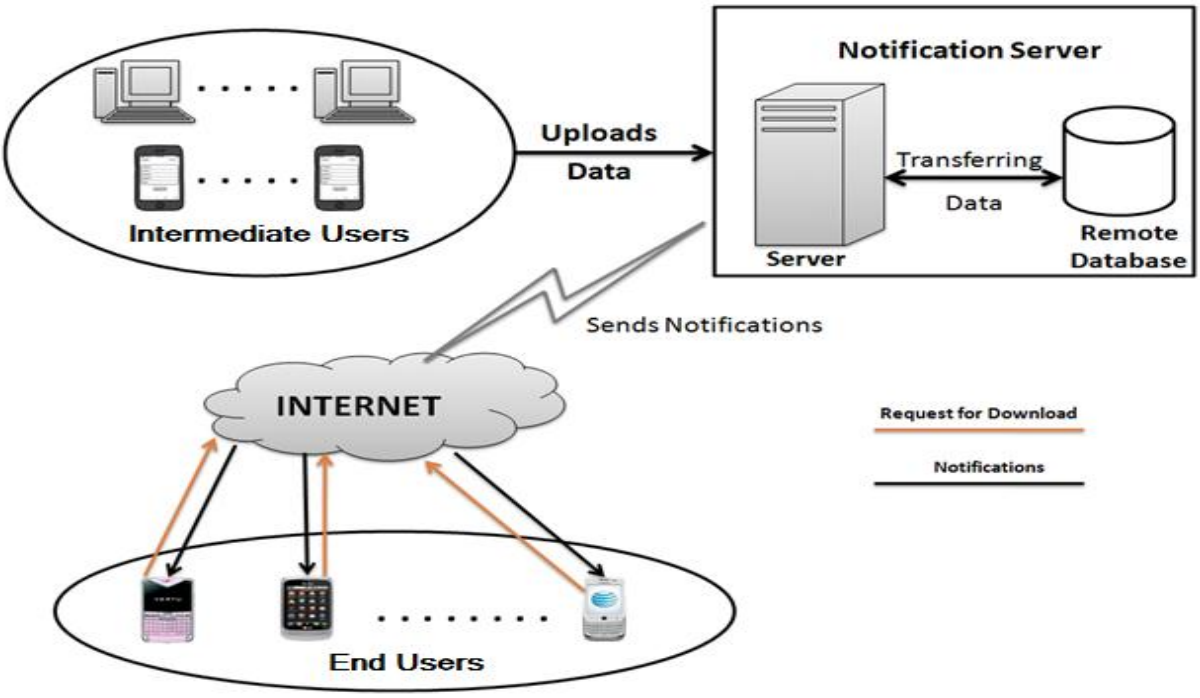


Fig. SYSTEM ARCHITECTURE

The system architecture consists of following main elements: a web server to host notices, an administrator to post notices, a server where notices will be uploaded and a remote database. The system has distinctive client-server architecture. There is a web server on one side and users on the client side. Every component has to play its own role independent task. The elements help the system work in systematic way and phases. The separate modules help the architecture to be efficient. The server is component where actually the notices are uploaded. The server is then responsible for forwarding the notices to intended users. Either group of users is informed about the notice or all users. The user who have Smartphones receive message on their phone's app. Other users who have ordinary phones will receive notice through text message. The Intermediate users are on who are provided with some authorities and privileges. They can initiate to send a notice. It is sent or uploaded on server and further server will work accordingly. The database is used to store user details. It also stores the notices that have been posted using system. The records of groups and privileges of the users are stored in database.

IV. SYSTEM IMPLEMENTATION PLAN

The system is mainly divided into two major components viz Web Portal and Android app.

A. *Development of Web Portal*

a) Content Management

The software for content management system is used. This will help to achieve various features for the web portal. We can manage large amount of data that stored and retrieved by many users. We can essentially assign privileges to different users. Using the content management we can manage the components and the web content efficiently.

b) User Management

User management is the feature that will help administrator to manage all the users. It also includes managing the filtering of users, controlling their login counts and login times. It will provide efficient authentication and authorization for accessing network resources. It will also provide secure control over state of users. We can do user synchronization and user mapping. A new user can be created , edited and they can be managed.

B. *Development of Android App*

An android app has to be developed for the end users. The app will provide efficient user interface. It will include all the functionalities like the user login and log out. The app will make use of the local database available on the phone for keeping the session track. The user if not logged in, then the notices will be send to his phone and will stored as unread. The notices can be verified by user as notices will provide the issue date and the valid date of the notice.

C. *Use of intermediate communicator*

A use of intermediate communicator will be done in order to transfer the data between other two modules. The exchange of information over the application can be done using intermediate communicator. This module will provide Internet communication between programs. The data and the content can be worked in this module. Basically this modules with the data transfer and management.

V. CONCLUSION

In this paper we focus on the analysis of the development of Instant Academic Messenger that will provide notices to each user. The usability of the system is very high, especially for its use within an institute. The system helps to communicate important messages to the end user instantly through the use of mobiles thus important messages reach the intended user at the right time.

The system can be extended for usage at the university level such as university examination form filling, seat allocation details etc. Inter- collegiate

communication can also be established via working in a collaborative environment with shared databases.

VI. **REFERENCES**

[1] PortableLab: Implementation of a Mobile Remote Laboratory for the Android Platform, Marco André Guerra, Instituto de Telecomunicações ,Lisboa, Portugal

mapg88@gmail.com, Cláudia Mariline Francisco, ESTSetúbal, IPS, Setúbal, Portugal

mariline.francisco@hotmail.com , Rui Neves Madeira, DSI, ESTSetúbal, IPS, Setúbal, Portugal, rui.madeira@estsetubal.ips.pt.

IT 30

Web Search Optimization By Mining Query Logs

Ashish Pathak
Dept. of computer engg.
PVG's COET,Pune
Pune, India
pathak.ashish3@gmail.com

Sonam Jain
Dept. of computer engg.
PVG's COET,Pune
Pune, India

Priyanka Chavan
Dept. of computer engg.
PVG's COET,Pune
Pune, India
priyankachavan88@gmail.com

Shashikant Gujrati
Dept. of computer engg.
PVG's COET,Pune
Pune, India

Abstract—The search engine retrieves relevant web pages from the index .The result is large set of web pages. To search for relevant web pages is very tedious job. Hence the web search results need to be optimized. Initially, web pages are displayed as a result from the index. On backend, the logs are maintained. The query log stores query words and clicked URLs for respective query words. The mathematical treatment computes the similarity values so as to form clusters i.e. potential group of queries. The sequential patterns of web pages are generated in each cluster. The online working rank updater module updates rank of web pages for display. The proposed system reduces user search space to improve ranks of desired web pages.

Keywords—Query Logs,Clustering,Sequential Pattern generator, Rank , Mining

Introduction

The world wide web is increasing day by day, and loads of pages are available on the web. This increase in the web has degraded the quality of search provided by search engines. This problem also arises due to the inability of the user to form

proper queries for the search engine. The user is many times not provided by the results that are very relevant and actually needed by him. The click through data of the user can be used to optimize results and provide more relevant search results. This paper proposes a method to cluster the user queries and click-data which update the rank of pages in the search result. The paper is organized as section II contains the related work, section III consists of the proposed method and finally conclusions and future work are given in section IV.

Related work

Web usage mining has been a subject of interest for intelligent web-based system. Data based mining has gained importance and it is making use of clickthrough data. Analysis of such query logs have been used by many researchers to provide personalization, to classify queries, to improve searching and for providing context during search.

In year 2002, a system was proposed to automatically optimize the retrieval quality of search engines. It makes use of Support Vector Machine approach for learning retrieval function.

In year 2009, an idea was proposed to make use of click through data to provide *Query Suggestion* technique that aimed to recommend relevant queries which potentially suit the information needs of that user. It then developed the novel and two level suggestion model by mining click through data. It was all done in the form of bipartite graph to provide semantically relevant queries.

One of the important task in this area is mining sequential patterns of web pages. AprioriAll algorithm which is three phase algorithm was used to find those patterns. Generalized Sequential Pattern is faster than AprioriAll algorithm, but SPADE is the best than all.

A critical look at the available literature indicates, search engines are using some sort of optimization measures on their search results but user is still posed to problems of finding the required information within the search result. The proposed method gives importance to information needs of users and reduces search space for user.

Our Sysytem

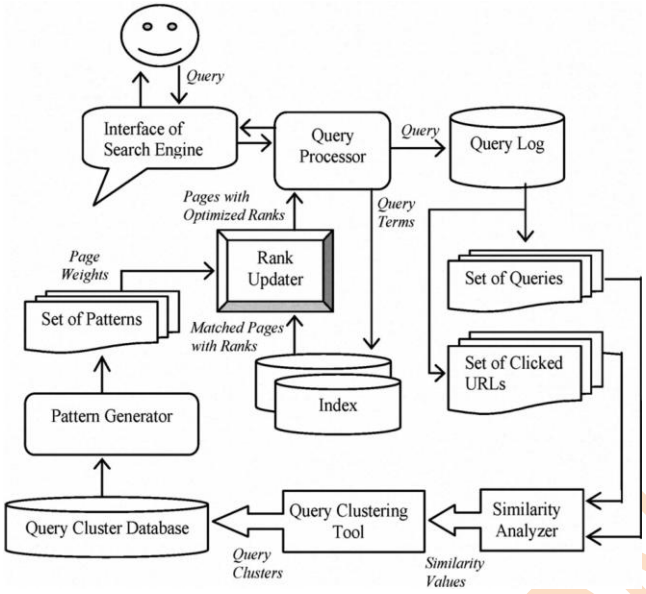
The main subparts of the system are –

A. Similarity Analyser

B. Query Clustering Tool

C. Sequential Pattern Generator

D. Rank Updater



Similarity Analyser

The analyzer is based on two criteria, query words based and clicked URLs based –

1) The keyword-based similarity function :

$$Simkeyword(p,q) = \frac{KW(p,q)}{\max(kw(p),kw(q))};$$

where kw( ) is the number of keywords in a query, KW(p, q) is the number of common keywords in two queries.

For example, the two queries “history of China” and “history of the United States” are very close queries (both asking about the history of a country).Their similarity is 0.33 on the basis of keywords.

2) The clicks-based similarity function:

$$Simclick(p,q) = \frac{RD(p,q)}{\max(rd(p),rd(q))}$$

where rd(p) and rd(q) are the number of referred documents for two queries p and q respectively, RD(p,q) is the number of document clicks in common.

3)The combined measure:

The two method above have its own advantages, but to make efficient grouping the combined similarity value is calculated.

$$Simcombined(p,q) = \alpha * Simkeyword(p,q) + \beta * Simclick(p,q);$$

Where α+β=1.For convenience both are taken as 0.5.

Query Clustering tool

Clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Precisely, data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses is to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available.

There are many clustering methods available, and each of them may give a different grouping of a dataset. Good clustering should have high intra cluster similarity and low inter cluster similarity. 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- Hierarchical methods and partitioning methods.

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 partitioning method, each object is a member of the cluster with which it is most similar; however the threshold of similarity has to be defined.

Here we will be using partition algorithm for clustering.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centre for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centre, 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 centre; 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.

Algorithm: Query_Clustering(Q,α,β,τ)

Given: A set of n queries and corresponding clicked URLs stored in an array Q[qi,URL1,..URLm],1≤i≤n

α=β=0.5;

Minimum similarity threshold τ;

Output: A set C= {C1, C2, … Ck} of k query clusters

// Start of Algorithm

//Let n be the number of documents to be clustered, k be the //no. of clusters.

findInitialSeeds (top N docs, k) {

create a clusterVector containing randomly selected J kn docs from top n docs

while(clusterVector.size > k){

maxSim = 0

for i = 1to k{

for j = i + 1 to k{

if sim(clusterVector[i], clusterVector[j]) > maxSim{

cluster1 = clusterVector[i]

cluster2 = clusterVector[j]

maxSim = sim(clusterVecor[i], clusterVector[j])

}

}

}

remove cluster1 and cluster2 from clusterVector

combine docs of cluster1 and cluster2 into cluster

add cluster to clusterVector

}

Return clusterVector

}

Where sim(clusterVector[i],clusterVector[j]) is a simcombined(p,q) calculated from following formulae -

$$Simkeyword(p,q) = \frac{KW(p,q)}{\max(kw(p),kw(q))};$$

$$Simclick(p,q) = \frac{RD(p,q)}{\max(rd(p),rd(q))};$$

$$Simcombined(p,q) = \alpha * Simkeyword(p,q) + \beta * Simclick(p,q);$$

Sequential Pattern Generator

Sequential patterns are an important issue in web usage mining. These patterns help in forming basis of intelligent web-based systems as it captures web browsing behavior of the user.

Capturing frequent user navigation patterns is very important in log mining. These patterns must be interpreted and knowledge must be extracted from them.

Here we are using SPADE algorithm to find frequent sequential patterns.

Two pre-requisite steps for this are:

1. Create Session id-timestamp list

The basic structure for this purpose is *Session id- timestamp list*. Session id timestamp list is a list which keeps session id and timestamp information for any patterns in all sessions.

After data preprocessing step we obtain a series of web pages visited in each session. Session can be given in following format :

$S_i = Page_{i-1} \rightarrow Page_{i-2} \rightarrow \ldots\ldots\ldots Page_{i-k} \rightarrow \ldots\ldots\ldots\ldots\ldots \rightarrow Page_{i-|Si|}$

The number k is the number of pages in the corresponding session. It can be used as a timestamp as follows.

Any pattern P_j is visited before P_k in session Sn if $j < k$. The timestamp information keeps the the order of last atom for patterns with length > 1. It keeps the order of atoms with length=1.

2. Construction of Pattern Lattice

Atom: Each single webpage in web domain called atom.

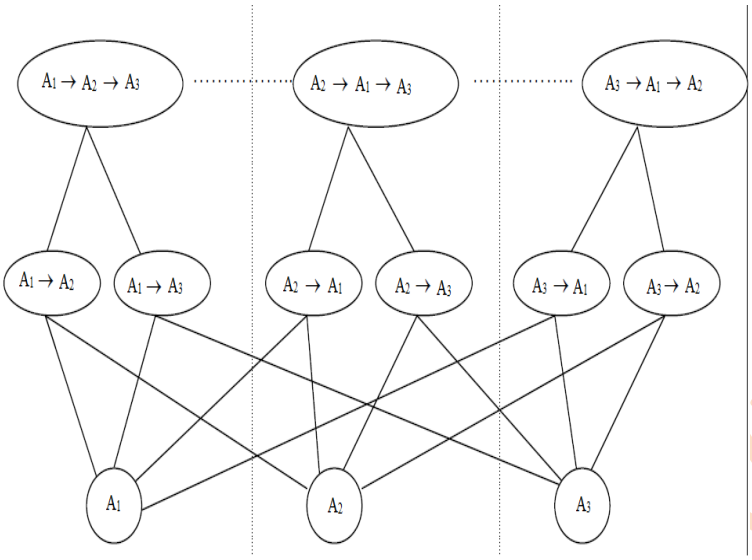
Pattern: Any Pattern consists of a sequential list of atoms.

Lattice: An algebra (L;v;∧) is called a lattice, if L is a nonempty set and v and ∧ are binary operations on L.

The bottom elements of patterns lattice are composed of single, as the basic element of pattern is atom. Each single atom stands for the length-1 prefix equivalence class. Beginning from bottom elements the frequency of upper elements with length *n* can be calculated by using two n-1 length patterns belonging to the same class.

Example :

Consider A_1, A_2 , A_3 be the atoms. Then the 3-length frequent pattern generation is as follows :



SPADE Algorithm: In the SPADE algorithm, firstly *Session id-timestamp list* of atoms created. Then these lists are sorted with respect to the support of each atom. Then, if support of any atom is below the input threshold it is eliminated automatically. Next, frequent patterns from single atoms are generated

according to union operation ∨ based on prefix-based approach defined above. Finally, all frequent items with length n > 2 are discovered in their length-1 prefix class [Atom] independently. While exploring search space both depth first and breadth first search can be used, however breadth first search has the advantage of keeping only current length-k pattern in the memory.

Algorithm:

SPADE(min_sup,S)

P1 = all frequent atoms with their Session id-timestamp list

P2 = all frequent length-2 patterns with their Session id-timestamp list

E = All equivalance classes of Atoms ∈ P1

For all [X] ∈ E do

Construct pattern lattice of [X]

Explore frequent patterns in [X] by

Using either Depth first or Breadth first search.

End For

End function

Rank Updater

In this module the rank of the pages that are returned by the query results are modified according to the user needs. The steps are followed as:

1. The cluster to which query belongs is found by similarity analyzer.
2. Sequential patterns are generated using the cluster to which query belongs.
3. The Weight of pages common in the sequential pattern and query results is calculated.
4. The re-rank is found by adding rank and weight of the page.

Weight(A)= $\ln(dep_{pat(A)})/depth(A)$;

Where $dep_{pat(A)}$ is depth of the pattern in which A belongs and depth(A) is depth of page A in pattern.

Rank Change of desired pages-

Re_Rank= Rank(A)+Weight(A);

IV Conclusion and future work

In the proposed approach we are collecting query words(keywords) and clicked URL by maintaining querylogs for user browsing session. On the basis of collected data from query logs we are making clusters and using clusters we are improving the ranks of pages. The returned pages with improved page ranks are directly mapped to the user feedbacks and dictate higher relevance than pages that exist in the result list but are never accessed by the user.

We can increase the efficiency of clustering algorithm by using buckshot algorithm also Bipartite graph techniques can be employed on query logs to retrieve a better clustering of user queries and thus returning more efficient results.

References

[11] Thorsten Joachirns, "Optimizing search engines using c1ickthrough data". Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data

[12] Edgar Meij, Marc Bron, Bouke Huurnink, Laura Hollink, and Maarten de Rijke, "Learning semantic query suggestions". In 8th International Semantic Web Conference (ISWC 2009), Springer, October 2009.

[13] Murat Ali Bayir, Ismail H. Toroslu, Ahmet Cosar, "A Performance Comparison of Pattern Discovery Methods on Web Log Data". Proceedings of 4th ACSIIEEE International Conference on Computer Systems and Applications (AICCSA-06)", pp. 445-451, 2006.

[14] A. Arasu, 1. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan, "Searching the Web," ACM Transactions on Internet Technology, Vol. I, No. I, pp. 97-101, 2001.

[15] J. Wen, J. Mie, and H. Zhang, "Clustering user queries of a search engine". In Proc. of 10th International World Wide Web Conference. W3C, 2001.

IT 031

IMAGE IDENTIFICATION USING CBIR

Suvarna Vitthal Khandagale, Kamthe Sweety Haridas , Salunke Rajashri R
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id-suvarnakhandagale30@gmail.com

Contact no.9730462558

*Annu Shabbir Tamboli
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id-tamboliannu@yahoo.in

Contact no.9096233439

Abstract

There is growing interest in CBIR because of the limitations inherent in metadata-based systems, as well as the large range of possible uses for efficient image retrieval. Textual information about images can be easily searched using existing technology, but requires humans to personally describe every image in the database. This is impractical for very large databases, or for images that are generated automatically, e.g. from surveillance cameras. It is also possible to miss images that use different synonyms in their descriptions. Systems based on categorizing images in semantic classes like "cat" as a subclass of "animal" avoid this problem but still face the same scaling issues. Our aim is to to build an application depending on Content-based image retrieval (CBIR). Our main aim is to filter the images and to retrieve the images that contains the data as per the query provided to the application. Secondary aim is to have this application utilized in law enforcement regarding access to the images. For examples we could make use of this application to keep a check on which images should be accessible to small children.

Keywords-CBIR, Color Histogram, Image retrieval, NNS, SIFT.

V. INTRODUCTION

The purpose of the project is to develop a CBIR based applications that could help in the filtering the images so as to get almost 100% accurate search results. Also using this application we could have law enforced for having access to the images.

Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the

application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases.

"Content-based" means that the search will analyze the actual contents of the image. The term 'content' in this context might refer colors, shapes, textures, or any other information that can be derived from the image itself. Without the ability to examine image content, searches must rely on metadata such as captions or keywords, which may be laborious or expensive to produce.

Innovativeness:T ill now we have search engines or application that perform search operation based on the query we send to it as input and match it with the names of the entity or the text. So we are able to extract only those entities which have proper naming i.e. name as per data or entity the file contains. With this application we are able to extract entities or files based on the data it actually contains. So, we do not have to depend on the naming system to extract exact data. This would more effective in case of searching images as they are the ones which difficult to extract if not properly named.

VI. TECHNIQUES USED IN CBIR

CBIR operates on a totally different principle, retrieving/searching stored images from a collection by comparing features automatically extracted from the images themselves. The commonest features used are mathematical measures of color, texture or shape (basic). A system (CBIR) allows users to formulate queries by submitting an example of the type of image being sought (input), though some offer alternatives such as selection from a palette or sketch input we can also select color textures or any other visual information. The system then identifies those stored images whose feature values match those of the query most closely (right side), and displays thumbnails of these images on the screen.

A. Colour

One of the most important features that make possible the recognition of images by humans is colour. Colour is a property that depends on the reflection of light to the eye and the processing of that information in the brain. We use colour everyday to tell the difference between objects, places, and the time of day [7]. Usually colours are defined in three dimensional colour spaces. These could either be ***RGB*** (Red, Green, and Blue), ***HSV*** (Hue, Saturation, and Value) or ***HSB*** (Hue, Saturation, and Brightness). The last two are dependent on the human perception of hue, saturation, and brightness.

Most image formats such as ***JPEG***, ***BMP***, ***GIF***, use the RGB colour space to store information [7]. The RGB colour space is defined as a unit cube with red, green, and blue axes. Thus, a vector with three co-ordinates represents the colour in this space. When all three coordinates are set to zero the colour perceived is black. When all three coordinates are set to 1 the colour perceived is white [7]. The other colour spaces operate in a similar fashion but with a different perception.

Methods of Representation

The main method of representing colour information of images in CBIR systems is through colour histograms. A colour histogram is a type of bar graph, where each bar represents a particular colour of the colour space being used. In MatLab for example you can get a colour histogram of an image in the RGB or HSV colour space. The bars in a colour histogram are referred to as bins and they represent the x-axis. The number of bins depends on the number of colours there are in an image. The y-axis denotes the number of pixels there are in each bin. In other words how many pixels in an image are of a particular colour.

An example of a color histogram in the HSV color space can be seen with the following image:

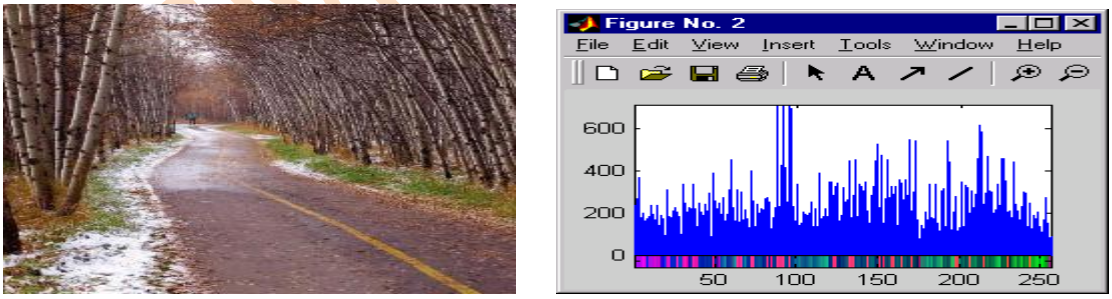


Fig. 1 Sample Image and its Corresponding Histogram

To view a histogram numerically one has to look at the color map or the numeric representation of each bin.

As one can see from the color map each row represents the color of a bin. The row is composed of the three coordinates of the color space. The first coordinate represents hue, the second saturation, and the third, value, thereby giving HSV. The percentages of each of these coordinates are what make up the color of a bin. Also one can see the corresponding pixel numbers for each bin, which are denoted by the blue lines in the histogram.

Quantization in terms of color histograms refers to the process of reducing the number of bins by taking colors that are very similar to each other and putting them in the same bin. By default the maximum number of bins one can obtain using the histogram function in MatLab is 256. For the purpose of saving time when trying to compare color histograms, one can quantize the number of bins. Obviously quantization reduces the information regarding the content of images but as was mentioned this is the tradeoff when one wants to reduce processing time.

Color Map (x-axis)			***Number of Pixels per Bin (y-axis)***
H	***S***	***V***	
0.9922	0.9882	0.9961	106
0.9569	0.9569	0.9882	242
0.9725	0.9647	0.9765	273
0.9176	0.9137	0.9569	372
0.9098	0.8980	0.9176	185
0.9569	0.9255	0.9412	204
0.9020	0.8627	0.8980	135
0.9020	0.8431	0.8510	166
0.9098	0.8196	0.8078	179
0.8549	0.8510	0.8941	188
0.8235	0.8235	0.8941	241
0.8471	0.8353	0.8549	104
0.8353	0.7961	0.8392	198

.	.	.	.
.	.	.	.
.	.	.	.

TABLE I COLOR MAP AND NUMBER OF PIXELS FOR THE PREVIOUS IMAGE

There are two types of colour histograms, Global colour histograms (***GCH***s) and Local colour histograms (***LCH***s). A GCH represents one whole image with a single colour histogram. An LCH divides an image into fixed blocks and takes the colour histogram of each of those blocks [7]. LCHs contain more information about an image but are computationally expensive when comparing images. "The GCH is the traditional method for colour based image retrieval. However, it does not include information concerning the colour distribution of the regions [7]" of an image. Thus when comparing GCHs one might not always get a proper result in terms of similarity of images.

B. Texture

Texture is that innate property of all surfaces that describes visual patterns, each having properties of homogeneity. It contains important information about the structural arrangement of the surface, such as; clouds, leaves, bricks, fabric, etc. It also describes the relationship of the surface to the surrounding environment [2]. In short, it is a feature that describes the distinctive physical composition of a surface.

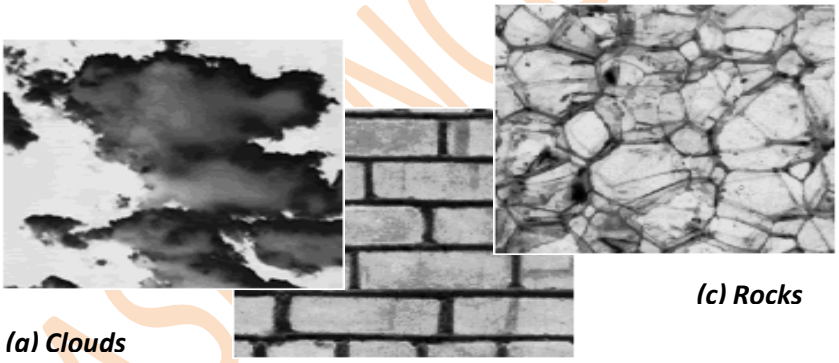


Fig. 2 Examples of textures

Texture properties include:

- Coarseness
- Contrast

- Directionality
- Line-likeness
- Regularity
- Roughness

Texture is one of the most important defining features of an image. It is characterized by the spatial distribution of gray levels in a neighborhood [8]. In order to capture the spatial dependence of gray-level values, which contribute to the perception of texture, a two-dimensional dependence texture analysis matrix is taken into consideration. This two-dimensional matrix is obtained by decoding the image file; jpeg, bmp, etc.

Methods of Representation

There are three principal approaches used to describe texture; statistical, structural and spectral.

- Statistical techniques characterize textures using the statistical properties of the grey levels of the points/pixels comprising a surface image. Typically, these properties are computed using: the grey level co-occurrence matrix of the surface, or the wavelet transformation of the surface.
- Structural techniques characterize textures as being composed of simple primitive structures called “texels” (or texture elements). These are arranged regularly on a surface according to some surface arrangement rules.
- Spectral techniques are based on properties of the Fourier spectrum and describe global periodicity of the grey levels of a surface by identifying high-energy peaks in the Fourier spectrum[9] .

For optimum classification purposes, what concern us are the statistical techniques of characterization… This is because it is these techniques that result in computing texture properties… The most popular statistical representations of texture are:

- Co-occurrence Matrix
- Tamura Texture
- Wavelet Transform

1) Co-occurrence Matrix

Originally proposed by R.M. Haralick, the co-occurrence matrix representation of texture features explores the grey level spatial dependence of texture [2]. A mathematical definition of the co-occurrence matrix is as follows [4]:

- Given a position operator ***P(i,j)***,
- let ***A*** be an n x n matrix

- Whose element ***A[i][j]*** is the number of times that points with grey level (intensity) ***g[i]*** occur, in the position specified by ***P***, relative to points with grey level ***g[j]***.
- Let ***C*** be the ***n x n*** matrix that is produced by dividing ***A*** with the total number of point pairs that satisfy ***P***. ***C[i][j]*** is a measure of the joint probability that a pair of points satisfying ***P*** will have values ***g[i]***, ***g[j]***.
- ***C*** is called a co-occurrence matrix defined by ***P***.

Examples for the operator ***P*** are: "***i*** above ***j***", or "***i*** one position to the right and two below ***j***", etc.

This can also be illustrated as follows… Let *t* be a translation, then a co-occurrence matrix C_t of a region is defined for every grey-level *(a, b)* by :

$$C_t(a,b) = card\{(s, s+t) \in R^2 | A[s] = a, A[s+t] = b\}$$

Here, $C_t(a, b)$ is the number of site-couples, denoted by *(s, s + t)* that are separated by a translation vector *t*, with *a* being the grey-level of *s*, and *b* being the grey-level of *s + t*.

For example; with an 8 grey-level image representation and a vector *t* that considers only one neighbor, we would find:

1	2	1	3	4
2	3	1	2	4
3	3	2	1	1

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	0	0	0	0	0
2	0	1	0	2	0	0	0	0
3	0	0	1	1	0	0	0	0
4	0	1	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

Fig. 3 Classical Co-occurrence matrix

At first the co-occurrence matrix is constructed, based on the orientation and distance between image pixels[2]. Then meaningful statistics are extracted from the matrix as the texture representation. Haralick proposed the different texture features[10].For each Haralick texture feature, we obtain a co-occurrence matrix. These co-occurrence matrices represent the spatial distribution and the dependence of the grey levels within a local area. Each ***(i,j)*** th entry in the matrices, represents the probability of going from one pixel with a grey level of '***i***'

to another with a grey level of '***j***' under a predefined distance and angle. From these matrices, sets of statistical measures are computed, called feature vectors[11] .

2) Tamura Texture

By observing psychological studies in the human visual perception, Tamura explored the texture representation using computational approximations to the three main texture features of: coarseness, contrast, and directionality[2,12]. Each of these texture features are approximately computed using algorithms...

- *Coarseness* is the measure of granularity of an image[12] , or average size of regions that have the same intensity [13].
- *Contrast* is the measure of vividness of the texture pattern. Therefore, the bigger the blocks that makes up the image, the higher the contrast. It is affected by the use of varying black and white intensities[12].
- *Directionality* is the measure of directions of the grey values within the image[12] .

3) Wavelet Transform

Textures can be modeled as quasi-periodic patterns with spatial/frequency representation. The wavelet transform transforms the image into a multi-scale representation with both spatial and frequency characteristics. This allows for effective multi-scale image analysis with lower computational cost [2]. According to this transformation, a function, which can represent an image, a curve, signal etc., can be described in terms of a coarse level description in addition to others with details that range from broad to narrow scales [11].

Unlike the usage of sine functions to represent signals in Fourier transforms, in wavelet transform, we use functions known as wavelets. Wavelets are finite in time, yet the average value of a wavelet is zero [2]. In a sense, a wavelet is a waveform that is bounded in both frequency and duration. While the Fourier transform converts a signal into a continuous series of sine waves, each of which is of constant frequency and amplitude and of infinite duration, most real-world signals (such as music or images) have a finite duration and abrupt changes in frequency. This accounts for the efficiency of wavelet transforms. This is because wavelet transforms convert a signal into a series of wavelets, which can be stored more efficiently due to finite time, and can be constructed with rough edges, thereby better approximating real-world signals [14].

Examples of wavelets are Coiflet, Morlet, Mexican Hat, Haar and Daubechies. Of these, Haar is the simplest and most widely used, while Daubechies have fractal structures and are vital for current wavelet applications [2]. These two are outlined below:

Haar Wavelet

The Haar wavelet family is defined as [2]:

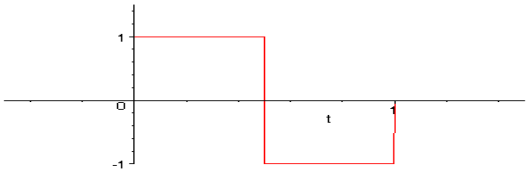


Fig. 4 Haar Wavelet Example

Daubechies Wavelet

The Daubechies wavelet family is defined as[2] :

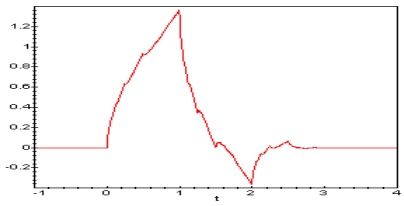


Fig. 5 Daubechies Wavelet Example

C. Shape[14]

Shape may be defined as the characteristic surface configuration of an object; an outline or contour. It permits an object to be distinguished from its surroundings by its outline[15] . Shape representations can be generally divided into two categories[2] :

- Boundary-based, and
- Region-based.

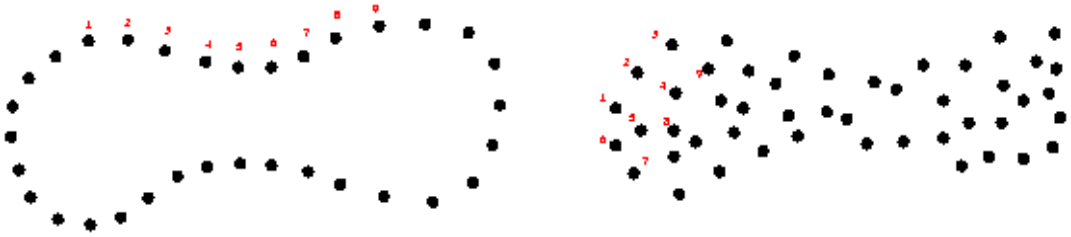


Fig. 6 Boundary-based & Region-based

Boundary-based shape representation only uses the outer boundary of the shape. This is done by describing the considered region using its external characteristics; i.e., the pixels along the object boundary. Region-based shape representation uses the entire shape region by describing the considered region using its internal characteristics; i.e., the pixels contained in that region [17].

Methods of Representation

For representing shape features mathematically, we have[16] :

Boundary-based:

- Polygonal Models, boundary partitioning
- Fourier Descriptors
- Splines, higher order constructs
- Curvature Models

Region-based:

- Super quadrics
- Fourier Descriptors
- Implicit Polynomials
- Blum's skeletons

The most successful representations for shape categories are Fourier Descriptor and Moment Invariants[2]:

- The main idea of Fourier Descriptor is to use the Fourier transformed boundary as the shape feature.
- The main idea of Moment invariants is to use region-based moments, which are invariant to transformations as the shape feature.

VII. ALGORITHMS

A. The SIFT (Scale-Invariant Feature Transform) algorithm[15]

The SIFT algorithm identifies features of an image that are distinct, and these features can in turn be used to identify similar or identical objects in other images. We will here give an introduction to the SIFT algorithm.

SIFT has four computational phases. The reason for this being that some computations performed by SIFT are very expensive. The cost of extracting the keypoints is minimized by the cascading approach of SIFT. The more expensive operations are only applied on locations that pass an initial, cheaper test. The output of the SIFT algorithm is a set of keypoint descriptors2. Once such descriptors have been generated for more than one image, one can begin image matching.

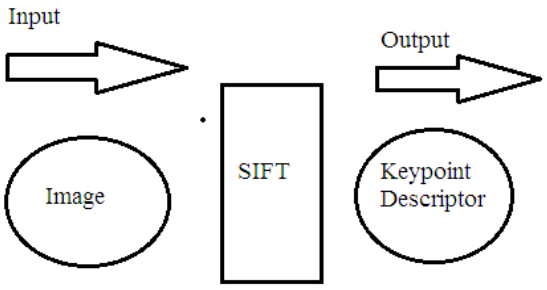


Fig. 7 Scale-Invariant Feature Transform

SIFT takes as input an image, and generates a set of keypoints descriptors. The keypoints descriptors may then be stored in a separate file.

The image matching, or object matching, is not part of the SIFT algorithm. For matching we use a nearest neighbor search (NNS), an algorithm that is able to detect similarities between keypoints. Thus, SIFT only makes matching possible by generating the keypoints descriptors.

B. Nearest neighbor search[15]

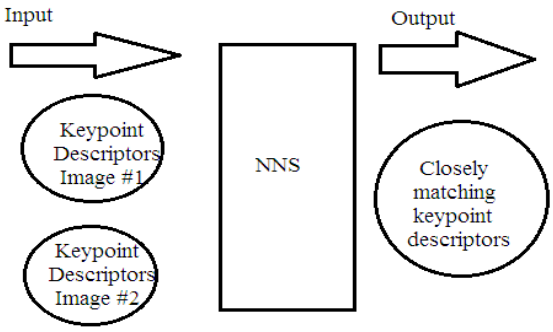


Fig. 8 Nearest neighbor search

When we check for an image match, the two sets of keypoints descriptors are given as input to a nearest neighbor search algorithm. The output of the algorithm is a set of keypoints descriptors found to be very similar.

Keypoints matching

As mentioned earlier, this is accomplished by a NNS in a kd-Tree. However, though the matching utilizes a NNS, the criterion for a match is not to be a nearest neighbor; then all nodes would have a match. Note, as stated before, the comparison is done from the source image, to the compared image. Thus, in the following outline, to "select a node" means to select a node from

the keypoints of the source image. The procedure is as follows:

1. Select a node from the set of all nodes not yet selected.

2. Mark the node as selected.

3. Locate the two nearest neighbors of the selected node.

4. If the distance between the two neighbors are less than or equal to a

given distance, we have a match. Mark the keypoints as match.

5. Perform step 1-4 for all nodes in the source image.

The key step of the matching algorithm is step 4. It is here it is decided whether the source node (or keypoints) has a match in the compared set of keypoints, or not.

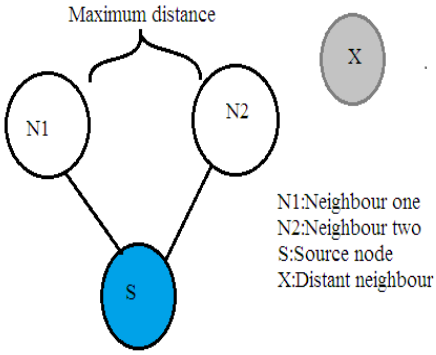


Fig. 9 Step of matching procedure

The image matching algorithm verifies the distance between the two closest neighbors of a source node. If the distance is within a given boundary, the source node is marked as a match.

Quality of Match formula[15]

Let *Ks* be the number of keypoints in the source image, *Kc* be the number of keypoints in the compared image, and *Km* be the number of matching keypoints.

We can rewrite the QoM formula to-

$$\mathbf{QoM}= \frac{Km*Kc}{Ks^{2}} * 100$$

Ks » Kc = very unreliable matching

Ks « Kc = very reliable matching

VIII. SYSTEM ARCHITECTURE

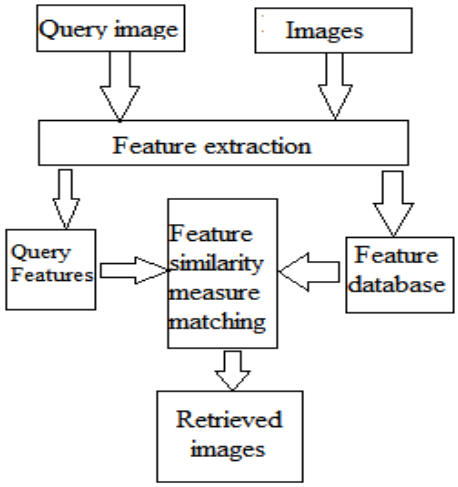


Fig. 10 Architecture of a typical CBIR system

For each image in the image database, its features are extracted and the obtained feature space (or vector) is stored in the feature database. When a query image comes in, its feature space will be compared with those in the feature database one by one and the similar images with the smallest feature distance will be retrieved.

IX. CONCLUSION

The application performs a simple colour-based search in an image database for an input query image, using colour histograms. It then compares the colour histograms of different images using the *Quadratic Distance Equation*. Further enhancing the search, the application performs a texture-based search in the colour results, using wavelet decomposition and energy level calculation. It then compares the texture features obtained using the *Euclidean Distance Equation*. A more detailed step would further enhance these texture results, using a shape-based search.

CBIR is still a developing science. As image compression, digital image processing, and image feature extraction techniques become more developed, CBIR maintains a steady pace of development in the research field. Furthermore, the development of powerful processing power, and faster and cheaper memories contribute heavily to CBIR development. This development promises an immense range of future applications using CBIR.

REFERENCES

[1] Barbeau Jerome, Vignes-Lebbe Regine, and Stamon Georges, "A Signature based on Delaunay Graph and Co-occurrence Matrix," Laboratoire Informatique et Systematique, Universiyt of Paris, Paris, France, July 2002, Found at:
a. http://www.math-info.univ-paris5.fr/sip-lab/barbeau/barbeau.pdf
[2] Sharmin Siddique, "A Wavelet Based Technique for Analysis and Classification of Texture Images," Carleton University, Ottawa, Canada, Proj. Rep. 70.593, April 2002.
[3] Thomas Seidl and Hans-Peter Kriegel, "Efficient User-Adaptable Similarity Search in Large Multimedia Databases," in Proceedings of the 23rd International Conference on Very Large Data Bases VLDB'97, Athens, Greece, August 1997, Found at:
a. http://www.vldb.org/conf/1997/P506.PDF
[4] FOLDOC, *Free On-Line Dictionary Of Computing*, "cooccurrence matrix," May 1995, [Online Document], Available at:

a. http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?cooccurrence+matrix
[5] Colin C. Venteres and Dr. Matthew Cooper, "A Review of Content-Based Image Retrieval Systems", [Online Document], Available at:
a. http://www.jtap.ac.uk/reports/htm/jtap-054.html
[6] Linda G. Shapiro, and George C. Stockman, *Computer Vision*, Prentice Hall, 2001.
[7] Shengjiu Wang, "A Robust CBIR Approach Using Local Color Histograms," Department of Computer Science, University of Alberta, Edmonton, Alberta, Canada, Tech. Rep. TR 01-13, October 2001, Found at:
a. http://citeseer.nj.nec.com/wang01robust.html
[8] R. Jain, R. Kasturi, and B. G. Schunck, *Machine Vision*, McGraw Hill International Editions, 1995.
[9] FOLDOC, *Free On-Line Dictionary Of Computing*, "texture," May 1995, [Online Document], Available at:
a. http://foldoc.doc.ic.ac.uk/foldoc/foldoc.cgi?query=texture
[10] "Texture," class notes for *Computerized Image Analysis MN2*, Centre for Image Analysis, Uppsala, Sweden, Winter 2002, Found at:
a. http://www.cb.uu.se/~ingela/Teaching/ImageAnalysis/Texture2002.pdf
[11] G. D. Magoulas, S. A. Karkanis, D. A. Karras and M. N. Vrahatis, "Comparison Study of Textural Descriptors for Training Neural Network Classifiers", in Proceedings of the 3rd IEEE-IMACS World Multi-conference on Circuits, Systems, Communications and Computers, vol. 1, 6221-6226, Athens, Greece, July 1999, Found at:
a. http://www.brunel.ac.uk/~csstgdm/622.pdf
[12] Pravi Techasith, "Image Search Engine," Imperial College, London, UK, Proj. Rep., July 2002, Found at:
a. http://km.doc.ic.ac.uk/pr-p.techasith-2002/Docs/OSE.doc
[13] Bjorn Johansson, "QBIC (Query By Image Content)", November 2002, [Online Document], Available at:
a. http://www.isy.liu.se/cvl/Projects/VISIT-bjojo/survey/surveyonCBIR/node26.html
[14] Content Based Image Retrieval By Rami Al-Tayeche (237262) & Ahmed Khalil (296918)
[15] An evaluation of the SIFT algorithm for CBIR by Thomas Bakken

IT 32

Performance Analysis of Reactive Routing Protocols for Mobile Ad Hoc Networks

Deepika A.Sarwate (9326972164)
deepika_medhekar@yahoo.com

Sheetal S.Patil (9420423491)
patilsheetal@yahoo.co.in

Abstract—Mobile Ad Hoc Network (MANET) is collection of multi-hop wireless mobile nodes that communicate with each other without centralized control or established infrastructure. The wireless links in this network are highly error prone and can go down frequently due to mobility of nodes, interference and less infrastructure. Therefore, routing in MANET is a critical task due to highly dynamic environment. In recent years, several routing protocols have been proposed for mobile ad hoc networks and prominent among them are DSR, AODV and TORA. This research paper provides an overview of these protocols by presenting their characteristics, functionality, benefits and limitations and then makes their comparative analysis so to analyze their performance. The objective is to make observations about how the performance of these protocols can be improved.

***Index Terms*—AODV, DSR, MANET, TORA**

I. INTRODUCTION

The wireless network can be classified into two types: Infrastructured or Infrastructure less. **In Infrastructured wireless** networks**,** the mobile node can move while communicating, the base stations are fixed and as the node goes out of the range of a base station, it gets into the range of another base station. The fig. 1, given below, depicts the Infrastructured wireless network.

In Infrastructureless or Ad Hoc wireless network, the mobile node can move while communicating, there are no fixed base stations and all the nodes in the network act as routers. The mobile nodes in the Ad Hoc network dynamically

establish routing among themselves to form their own network 'on the fly'. This type of network can be shown as in fig. 2.

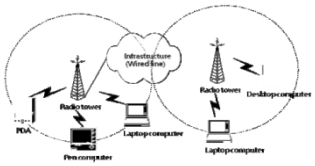


Fig. 1: Infrastructured Wireless Networks

Fig. 2: Infrastructureless or Ad Hoc Wireless Networks

A Mobile Ad Hoc Network (MANET) is a collection of wireless mobile nodes forming a temporary/short-lived network without any fixed infrastructure where all nodes are free to move about arbitrarily and where all the nodes configure themselves. In MANET, each node acts both as a router and as a host & even the topology of network may also change rapidly. Some of the challenges in MANET include:

1) Unicast routing
2) Multicast routing
3) Dynamic network topology
4) Speed
5) Frequency of updates or Network overhead
6) Scalability
7) Mobile agent based routing
8) Quality of Service
9) Energy efficient/Power aware routing
10) Secure routing

The key challenges faced at different layers of MANET are shown in Fig. 3. It represents layered structure and approach to ad hoc networks.

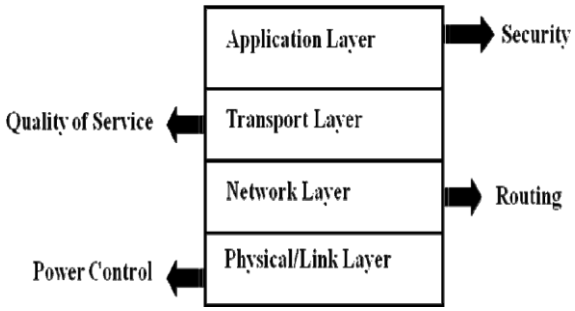


Fig.3: MANET Challenges

II. ROUTING PROTOCOLS

A routing protocol is needed whenever a packet needs to be transmitted to a destination via number of nodes and numerous routing protocols have been proposed for such kind of ad hoc networks. These protocols find a route for packet delivery and deliver the packet to the correct destination. The studies on various aspects of routing protocols have been an active area of research for many years. Many protocols have been suggested keeping applications and type of network in view. Basically, routing protocols can be broadly classified into two types as (a) Table Driven Protocols or Proactive Protocols and (b) On-Demand Protocols or Reactive Protocols

Table Driven or Proactive Protocols: In Table Driven routing protocols each node maintains one or more tables containing routing information to every other node in the network. All nodes keep on updating these tables to maintain latest view of the network. Some of the existing table driven or proactive protocols are: DSDV [6], [19], DBF [7], WRP [23] and ZRP [27], [13].

On Demand or Reactive Protocols: In these protocols, routes are created as and when required. When a transmission occurs from source to destination, it invokes the route discovery procedure. The route remains valid till destination

is achieved or until the route is no longer needed. Some of the existing on demand routing protocols are: DSR [8], [9], AODV [4], [5] and TORA [26], [27].

The emphasis in this research paper is concentrated on the survey and comparison of various On Demand/Reactive Protocols such as DSR, AODV and TORA as these are best suited for Ad Hoc Networks. The next sub-section describes the basic features of these protocols.

III. DYNAMIC SOURCE ROUTING [8, 9]

Dynamic Source Routing (DSR) is an Ad Hoc routing protocol which is based on the theory of source-based routing rather than table-based. This protocol is source-initiated rather than hop-by-hop. This is particularly designed for use in multi hop wireless ad hoc networks of mobile nodes. Basically, DSR protocol does not need any existing network infrastructure or administration and this allows the Network to be completely self-organizing and self-configuring. This Protocol is composed of two essential parts of route discovery and route maintenance. Every node maintains a cache to store recently discovered paths. When a node desires to send a packet to some node, it first checks its entry in the cache. If it is there, then it uses that path to transmit the packet and also attach its source address on the packet. If it is not there in the cache or the entry in cache is expired (because of long time idle), the sender broadcasts a route request packet to all of its neighbors asking for a path to the destination. The sender will be waiting till the route is discovered. During waiting time, the sender can perform other tasks such as sending/forwarding other packets. As the route request packet arrives to any of the nodes, they check from their neighbor or from their caches whether the destination asked is known or unknown. If route information is known, they send back a route reply packet to the destination otherwise they broadcast the same route request packet. When the route is discovered, the required packets will be transmitted by the sender on the discovered route. Also an entry in the cache will be inserted for the future use. The node will also maintain the age information of the entry so as to know whether the cache is fresh or not. When a data packet is received by any intermediate node, it first checks whether the packet is meant for itself or not. If it is meant for itself (i.e. the intermediate node is the destination), the packet is received otherwise the same will be forwarded using the path attached on the data packet. Since in Ad hoc network, any link might fail anytime. Therefore, route maintenance process will constantly monitors and will also notify the nodes if there is any failure in the path. Consequently, the nodes will change the entries of their route cache.

Benefits and Limitations of DSR

One of the main benefit of DSR protocol is that there is no need to keep routing table so as to route a given data packet as the entire route is contained in the packet header. The limitations of DSR protocol is that this is not scalable to large networks and even requires significantly more processing resources than most other protocols. Basically, In order to obtain the routing information, each node must spend lot of time to process any control data it receives, even if it is

not the intended recipient. The flowchart [17] for DSR Protocol is given below:

IV. ADOV (AD HOC ON DEMAND DISTANCE VECTOR) [4], [5]

AODV is a variation of Destination-Sequenced Distance-Vector (DSDV) routing protocol which is collectively based on DSDV and DSR. It aims to minimize the requirement of system-wide broadcasts to its extreme. It does not maintain routes from every node to every other node in the network rather they are discovered as and when needed & are maintained only as long as they are required.

The key steps of algorithm used by AODV for establishment of unicast routes are explained below.

A. Route Discovery

When a node wants to send a data packet to a destination node, the entries in route table are checked to ensure whether there is a current route to that destination node or not. If it is there, the data packet is forwarded to the appropriate next hop toward the destination. If it is not there, the route discovery process is initiated. AODV initiates a route discovery process using Route Request (RREQ) and Route Reply (RREP). The source node will create a RREQ packet containing its IP address, its current sequence number, the destination's IP address, the destination's last sequence number and broadcast ID. The broadcast ID is incremented each time the source node initiates RREQ. Basically, the sequence numbers are used to determine the timeliness of each data packet and the broadcast ID & the IP address together form a unique identifier for RREQ so as to uniquely identify each request. The requests are sent using RREQ message and the information in connection with creation of a route is sent back in RREP message. The source node broadcasts the RREQ packet to its neighbours and then sets a timer to wait for a reply. To process the RREQ, the node sets up a reverse route entry for the source node in its route table. This helps to know how to forward a RREP to the source. Basically a lifetime is associated with the reverse route entry and if this entry is not used within this lifetime, the route information is deleted. If the RREQ is lost during transmission, the source node is allowed to broadcast again using route discovery mechanism.

B. Expanding Ring Search Technique

The source node broadcasts the RREQ packet to its neighbours which in

turn forwards the same to their neighbours and so forth. Especially, in case of large network, there is a need to control network-wide broadcasts of RREQ and to control the same; the source node uses an expanding ring search technique. In this technique, the source node sets the Time to Live (TTL) value of the RREQ to an initial start value. If there is no reply within the discovery period, the next RREQ is broadcasted with a TTL value increased by an increment value. The process of incrementing TTL value continues until a threshold value is reached, after which the RREQ is broadcasted across the entire network.

C. Setting up of Forward Path

When the destination node or an intermediate node with a route to the destination receives the RREQ, it creates the RREP and unicast the same towards the source node using the node from which it received the RREQ as the next hop. When RREP is routed back along the reverse path and received by an intermediate node, it sets up a forward path entry to the destination in

its routing table. When the RREP reaches the source node, it means a route from source to the destination has been established and the source node can begin the data transmission.

D. Route Maintenance

A route discovered between a source node and destination node is maintained as long as needed by the source node. Since there is movement of nodes in mobile ad hoc network and if the source node moves during an active session, it can reinitiate route discovery mechanism to establish a new route to destination.

Conversely, if the destination node or some intermediate node moves, the node upstream of the break initiates Route Error (RERR) message to the affected active upstream neighbors/nodes. Consequently, these nodes propagate the RERR to their predecessor nodes. This process continues until the source node is reached. When RERR is received by the source node, it can either stop sending the data or reinitiate the route discovery mechanism by sending a new RREQ message if the route is still required.

E. Benefits and Limitations of AODV

The benefits of AODV protocol are that it favors the least congested route instead of the shortest route and it also supports both unicast and multicast packet transmissions even for nodes in constant movement. It also responds very quickly to the topological changes that affects the active routes. AODV does not put any additional overheads on data packets as it does not make use of source routing.

The limitation of AODV protocol is that it expects/requires that the nodes in the broadcast medium can detect each others' broadcasts. It is also possible that a valid route is expired and the determination of a reasonable expiry time is difficult. The reason behind this is that the nodes are mobile and their sending rates may differ widely and can change dynamically from node to node. In addition, as the size of network grows, various performance metrics begin decreasing. AODV is vulnerable to various kinds of attacks as it based on the assumption that all nodes must cooperate and without their cooperation no route can be established.

V. TORA (TEMPORARY ORDERED ROUTING PROTOCOL) [26], [27]

TORA is a distributed highly adaptive routing protocol designed to operate in a dynamic multihop network. TORA uses an arbitrary height parameter to determine the direction of link between any two nodes for a given destination. Consequently, multiple routes often exist for a given destination but none of them are necessarily the shortest route. To initiate a route, the node broadcasts a QUERY packet to its neighbors. This QUERY is rebroadcasted through the network until it reaches the destination or an intermediate node that has a route to the destination. The recipient of the QUERY packet then broadcasts the UPDATE packet which lists its height with respect to the destination. When this packet propagates in the network, each node that receives the UPDATE packet sets its height to a value greater than the height of the neighbour from which the UPDATE was received. This has the effect of creating a series of directed links from the original sender of the QUERY packet to the node that initially generated the UPDATE packet. When it was discovered by a node that the route to a destination is no longer valid, it will

adjust its height so that it will be a local maximum with respect to its neighbours and then transmits an UPDATE packet. If the node has no neighbors of finite height with respect to the destination, then the node will attempt to discover a new route as described above. When a node detects a

network partition, it will generate a CLEAR packet that results in reset of routing over the ad hoc network. The flowchart [17] for TORA Protocol is given below:

A. Benefits and Limitations of TORA

One of the benefits of TORA is that the multiple routes between any source destination pair are supported by this protocol. Therefore, failure or removal of any of the nodes is quickly resolved without source intervention by switching to an alternate route.

TORA is also not free from limitations. One of them is that it depends on synchronized clocks among nodes in the ad hoc network. The dependence of this protocol on intermediate lower layers for certain functionality presumes that the link status sensing, neighbor discovery, in order packet delivery and address resolution are all readily available. The solution is to run the Internet MANET Encapsulation Protocol at the layer immediately below TORA. This will make the overhead for this protocol difficult to separate from that imposed by the lower layer.

B. 3.0 Performance Metrics

There are number of qualitative and quantitative metrics that can be used to compare reactive routing protocols. Most of the existing routing protocols ensure the qualitative metrics. Therefore, the following different quantitative metrics have been considered to make the comparative study of these routing protocols through simulation.

1) **Routing overhead:** This metric describes how many routing packets for route discovery and route maintenance need to be sent so as to propagate the data packets.

2) **Average Delay:** This metric represents average end-to-end delay and indicates how long it took for a packet to travel from the source to the application layer of the destination. It is measured in seconds.

3) **Throughput:** This metric represents the total number of bits forwarded to higher layers per second. It is measured in bps. It can also be defined as the total amount of data a receiver actually receives from sender divided by the time taken by the receiver to obtain the last packet.

4) **Media Access Delay:** The time a node takes to access media for starting the packet transmission is called as media access delay. The delay is recorded for

each packet when it is sent to the physical layer for the first time.

5) **Packet Delivery Ratio:** The ratio between the amount of incoming data packets and actually received data packets.

6) **Path optimality:** This metric can be defined as the difference between the path actually taken and the best possible path for a packet to reach its destination.

VI. CONCLUSION

In this research paper, an effort has been made to concentrate on the comparative study and performance analysis of various on demand/reactive routing protocols (DSR, AODV and TORA) on the basis of above mentioned performance metrics. The results after analysis have reflected in Table I and Table II. The first table is description of parameters selected with respect to low mobility and lower traffic. It has been observed that the performance of all protocols studied was almost stable in sparse medium with low traffic. TORA performs much better in packet delivery owing to selection of better routes using acyclic graph. Table II is evaluation of same parameters with increasing speed and providing more nodes. The results indicate that AODV keeps on improving with denser mediums and at faster speeds.

Table III is description of other important parameters that make a protocol robust and steady in most cases. The evaluation predicts that in spite of slightly more overhead in some cases DSR and AODV outperforms TORA in all cases. AODV is still better in Route updation and maintenance process.

It has been further concluded that due to the dynamically changing topology and infrastructure less, decentralized characteristics, security and power awareness is hard to achieve in mobile ad hoc networks. Hence, security and power awareness mechanisms should be built-in features for all sorts of applications based on ad hoc network. The focus of the study is on these issues in our future research work and effort will be made to propose a solution for routing in Ad Hoc networks by tackling these core issues of secure and power aware/energy efficient routing.

REFERENCES

[1] Sunil Taneja*and Ashwani Kush†, “A Survey of Routing Protocols in Mobile Ad HocNetworks International Journal of Innovation,

Management and Technology, Vol. 1, No. 3, August 2010

ISSN: 2010- 0248

[2] Anne Aaron, Jie Weng, "Performance Comparison of Ad-hoc Routing Protocols for Networks with Node Energy Constraints", available at http://ivms.stanford.edu

[3] Charles Perkins, Elizabeth Royer, Samir Das, Mahesh Marina, "Performance of two on-demand Routing Protocols for Ad-hoc Networks", IEEE Personal Communications, February 2001, pp.

16-28.

[4] C. Perkins, E. B. Royer, S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing - Internet Draft", RFC 3561, IETF Network Working Group, July 2003.

[5] C. E. Perkins and E. M. Royer, "Ad-Hoc On Demand Distance Vector Routing", Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications (WMCSA), New Orleans, LA,

1999, pp. 90-100.

[6] Perkins, Charles E. and Bhagwat, Pravin (1994).Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers.http://www.cs.virginia.edu/~cl7v/cs851-papers/dsdv-sigcomm94.pdf. Retrieved 2006-10-20.

[7] Andrew S. Tannenbaum: Computernet Network, 4th edition, Prentice Hall

[8] D. B. Johnson, D. A. Maltz, Y.C. Hu, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)", IETF Draft, April 2003, work in progress. http://www.ietf.org/internet-drafts/draft-ietf-manet-dsr-09.txt

[9] D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad Hoc Networks", Mobile Computing, T. Imielinski and H. Korth, Eds., Kulwer Publ., 1996, pp. 152-81.

[10] David A. Maltz, "On-Demand Routing in Multi-hop Wireless Mobile

Ad Hoc Networks', May 2001, available at www.monarch.cs.rice.edu

[11] E.M.Rover, C.K.Toh, "A review of current routing protocols for ad hoc networks", IEEE Communications, vol 6, 1999, pp 46-55.

[12] F. Bertocchi, P. Bergamo, G. Mazzin, "Performance Comparison of Routing Protocols for Ad hoc Networks", IEEE GLOBECOM 2003.

[13] Haas, Z. J., 1997 (ps). A new routing protocol for the reconfigurable wireless networks. Retrieved 2011-05-06.

[14] H. Ehsan and Z. A. Uzmi (2004), "Performance Comparison of Ad HocWireless Network Routing Protocols", IEE,E 8th International Multitopic Conference, Proceedingsof INMIC, December 2004, pp.457 – 465.

[15] Iskra Djonova Popova, "A PowerPoint presentation on Routing in Ad-hoc Networks", 9th CEENet Workshop on Network Technology, Budapest 2004.

[16] J. Broch, D.A. Maltz, D. B. Johnson, Y-C. Hu, J. Jetcheva, "A performance comparison of Multi-hop wireless ad-hoc networking routing protocols", in the proceedings of the 4th International Conference on Mobile Computing and Networking (ACM MOBICOM '98), October 1998, pages 85-97.

[17] Md. Golam Kaosar, Hafiz M. Asif, Tarek R. Sheltami, Ashraf S. Hasan Mahmoud, "Simulation-Based Comparative Study of On Demand Routing Protocols for MANET", available at http://www.lancs.ac.uk

[18] Per Johansson, Tony Larsson, Nicklas Hedman, Bartosz Mielczarek, "Routing protocols for mobile ad-hoc networks – a comparative performance analysis", in the proceedings of the 5th International Conference on Mobile Computing and Networking (ACM MOBICOM '99), August 1999, pages 195-206.

[19] P. Chenna Reddy, Dr. P. Chandrasekhar Reddy, "Performance Analysis of Adhoc Network Routing Protocols", Academic Open Internet Journal, SSN 1311-4360, Volume 17, 2006

[20] R. Misra, C. R. Manda, "Performance Comparison of AODV/DSR On-Demand Routing Protocols for Ad Hoc Networks in Constrained Situation", IEEE ICPWC 2005.

[21] S. Gowrishankar, T.G. Basavaraju, M. Singh, Subir Kumar Sarkar, "Scenario based Performance Analysis of AODV and OLSR in Mobile Ad hoc Networks", available at http://www.ijcim.th.org

[22] Samir R. Das, Charles E. Perkins, Elizabeth M. Royer, "Performance Comparison of Two On-demand Routing Protocols for Ad Hoc Networks", in the

proceedings of NFOCOM 2000, Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, volume 1, pages 3 – 12 and also available at www.cs.ucsb.edu

[23] Murthy, Shree; Garcia-Luna-Aceves, J. J. (1996-10-01), "An efficient routing protocol for wireless networks", Mobile Networks and Applications (Hingham, MA: Kluwer Academic Publishers) 1 (2): 183–197,doi:10.1007/BF01193336

[24] V. Nazari, K. Ziarati, "Performance Comparison of Routing Protocols for Mobile Ad hoc Networks", IEEE 2006.

[25] V. Park and S. Corson, Temporally Ordered Routing Algorithm (TORA) Version 1, Functional specification IETF Internet draft, http://www.ietf.org/internet-drafts/draft-ietf-manet-tora-spec-01.txt,

1998.

[26] V. D. Park and M. S. Corson, "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks", Proceedings of the IEEE International Conference on Computer Communications (INFOCOM), Kobe, Japan, 1997, pp. 1405-1413.

[27] Z. J. Hass and M. R. Pearlman, "Zone Routing Protocol (ZRP)", Internet draft available at www.ietf.org.

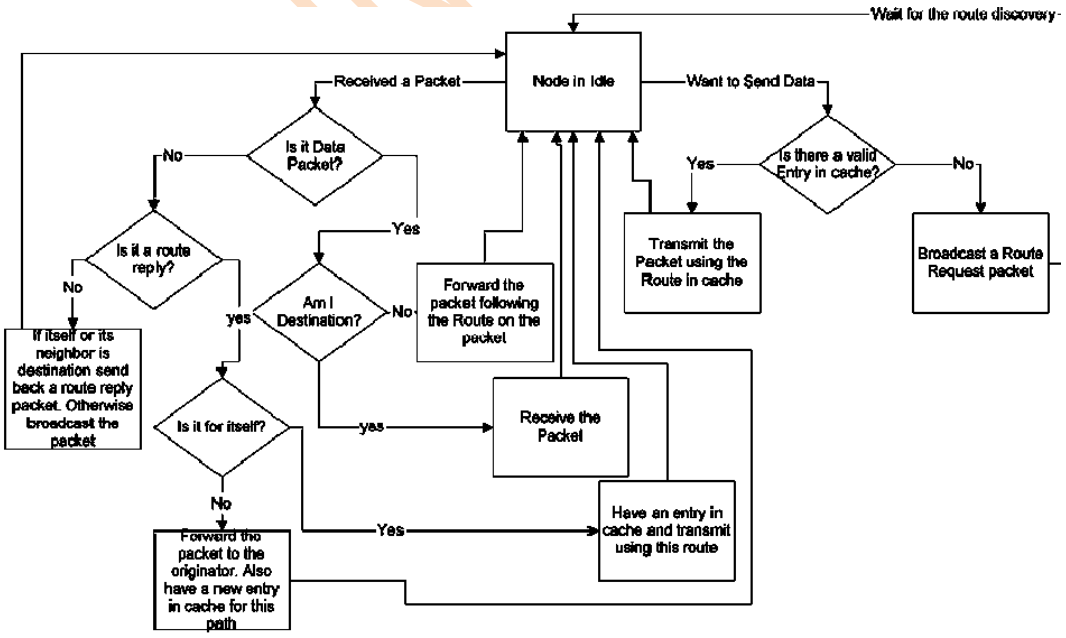


Fig. 4: Flow chart for DSR Working

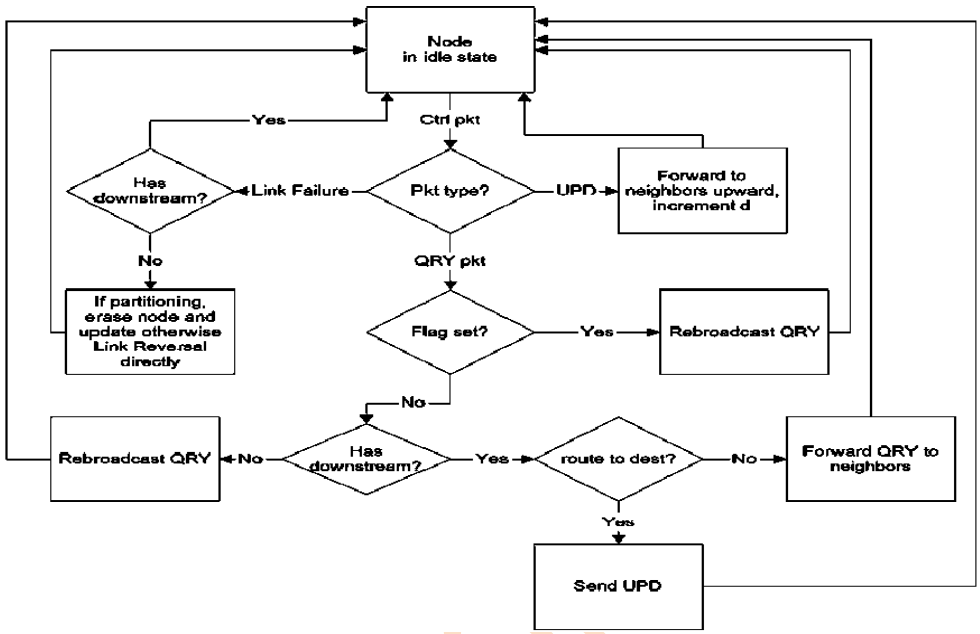


Fig. 5: Flow chart for TORA

TABLE I: METRICS W.R.T LOW MOBILITY

Low Mobility and Low Traffic				
Protocol	**Routing Overhead**	**Average end to end delay**	**Packet delivery**	**Path Optimality**
DSR	Low	Average	High	Average
AODV	Low	Average	High	Average

TORA	Moderate	Low	High	Good

TABLE II: METRICS W.R.T HIGH MOBILITY

High Mobility and Low Traffic				
Protocol	**Routing Overhead**	**Average end to end delay**	**Packet delivery**	**Path Optimality**
DSR	Average	Average	Average	Low
AODV	Very High	Average	Average	Average
TORA	High	More	Low	Average

TABLE III: EVALUATION W.R.T OTHER PARAMETERS

Protocol	**Category**	**Protocol Type**	**Loop Freedom**	**Multiple routes**	**Multicast**	**Security**	**Message Overhead**	**Periodic broadcast**	**Requires sequence data**	**Expiry of routing information**	**Summary**
DSR	On Demand	Source Routi	Yes	Yes	No	No	High	No	No	No	Route Discov
AODV	On Demand or Reactive	Distance Vector	Yes	No	Yes	No	High	Possible	Yes	Yes	Route Discovery, Expanding Ring Searc

TORA	On Demand or Reactive	Link Reversal	No	No	No	No	Moderate	Possible	Yes	No	Route UPDATE packets

IT 33

A Comparative study of selected security tools for personal computers

Mrs. Deepika A. Sarwate, (Lecturer) deepika_medhekar@yahoo.com
(9326972164))

JSPM Narhe Technical Campus, RSSCA,

Mrs Sheetal S.Patil(Lecturer) patilsheetal@yahoo.co.in
(9420423491)

JSPM Narhe Technical Campus,RSSCA,

Abstract—

It is difficult to imagine life without personal computers. While personal computers have software programs are available to prevent hackers from invading the information you have stored on your computer. People love surfing and shopping on the Internet and most do not realize how dangerous it can be. Even receiving and opening emails can be hazardous. It is important for users of personal computers to learn about software and other tools to protect their computers and their personal information.

There are number of venders providing antivirus packages which are differ in various features such as installation time, size, memory utilized, boot time, user interface launch time and full system scan time, open close document scan time, email scan time etc. Comparative study of antivirus tools is to be done in order to suggest the proper security tool depending upon the infrastructure need of the user.

Antivirus selection Management is a process in which one establishes and maintains policies, procedures, and practices required for protecting personal computer information system asset. The various tools & steps used today for maintaining corporate network security. The outcome of this work will help to various part of the society in deciding for what type of antivirus software they should go for according the need of infrastructure available.

Computer virus research is a fascinating subject to many who are interested in nature, biology, or mathematics. Everyone who uses a computer will likely encounter some form of the increasingly common problem of computer viruses.

In fact, some well-known computer virus researchers became interested in the field when, decades ago, their own systems were infected. [1].

This research paper highlights the performance of antivirus software using the number of parameters such as installation time, size, memory utilised, boot time, user interface launch time and full system scan time etc.

Keywords — Personal Computer (PC), Antivirus(AV), virus,Total security(TS), internet security(IS), Security, System, Installation Size(INS), installation tim(INT), memory utilization(MU),, boot time (BT), user interface launch time(UILT) and full scan time(FST)

1. Introduction

If there's one word that can strike fear in the heart of any computer user, especially one who accesses the internet, or exchanges diskettes, that word is, "virus." Viruses can generate so much fear in the cyber world that news of a new virus often spreads faster than the virus itself.

A **personal computer** (**PC**) is any general-purpose computer whose size, capabilities, and original sales price make it useful for individuals, and which is intended to be operated directly by an end-user with no intervening computer operator.While personal computers have software programs are available to prevent hackers from invading the information you have stored on your computer. People love surfing and shopping on the Internet and most do not realize how dangerous it can be. Even receiving and opening emails can be hazardous. It is important for users of personal computers to learn about software and other tools to protect their computers and their personal information.

In order to prevent such data looses many organization came forward and designed network security tools and antivirus packages. Antivirus packages are mainly used to prevent and remove the viruses, Trojans, worms etc, where as firewalls are used to monitor incoming and outgoing connections.

Computers are used extensively to process the data and to provide information for decision making therefore it is necessary to control its use. Due to organizational cost of data loss, cost of incorrect decision making, and value of computer software hardware organizations suffer a major loss therefore the integrity of data and information must be maintained.

Antivirus packages are mainly used to safeguard. There are number of venders providing antivirus packages which are differ in various features such as installation time, size, memory utilized, boot time, user interface launch time and full system scan time etc.

2. Virus and Antivirus Overview

A computer virus is self replicating program containing code that explicitly copies itself and that can infects other program by modifying them or their environment [2].Broadly, viruses can be found in following variations.

Trojan Horse:

A trojan horse program has the appearance of having a useful and desired function. While it may advertise its activity after launching, this information is not apparent to the user beforehand. Secretly the program performs other, undesired functions. A Trojan Horse neither replicates nor copies itself, but causes damage or compromises the security of the computer. A Trojan Horse must be sent by someone or carried by another program and may arrive in the form of a joke program (This is not a virus but a trick that aims to make users believe they have been infected by a virus) or software of some sort.

Worms:

A worm is a program that makes and facilitates the distribution of copies of itself; for example, from one disk drive to another, or by copying itself using email or another transport mechanism. The worm may do damage and compromise the security of the computer. It may arrive via exploitation of a system vulnerability or by clicking on an infected e-mail.
Bootsector Virus: A virus that infects the first few sectors of computer hard drive or diskette drive allowing the virus to activate as the driver or diskette boots. These are normally spread by floppy disks.

Macro Virus: Macro viruses are viruses that use another application's macro programming language to distribute themselves. They infect documents such as MS Word or MS Excel and are typically spread to other similar documents. They are the first viruses to infect data files, rather than executables. Data files, to which macros are attached, provide viruses with a more effective replication method than executable files. Data files are exchanged far more frequently than executable files. If you add the increased use of e-mail [and the ability to attach files to e-mail], and mass access to the Internet [and on-line services like

CompuServe and America Online], this is likely to make macro viruses a much greater threat to computer users than 'traditional' file viruses. Macro viruses are not platform-specific.

Overwriting viruses An overwriting virus simply overwrites each file it infects with itself, and the program no longer functions. Because this is so glaringly obvious, overwriting viruses are never successful in spreading.

Memory Resident Viruses: Memory Resident Viruses reside in a computers volitale memory (RAM). They are initiated from a virus which runs on the computer and they stay in memory

after it's initiating program closes.

Rootkit Virus: A rootkit virus is an undetectable virus which attempts to allow someone to gain control of a computer system. The term rootkit comes from the linux administrator root user. These viruses are usually installed by trojans and are normally disguised as operating system files.

Polymorphic Viruses: A polymorphic virus not only replicates itself by creating multiple files of itself, but it also changes it's digital signature every time it replicates. This makes it difficult for less sophisticated antivirus software to detect.

Logic Bombs/Time Bombs: These are viruses which are programmed to initiate at a specific date or when a specific event occurs. Some examples are a virus which deletes your photos on Halloween, or a virus which deletes a database table if a certain employee gets fired.

Spyware: Spyware is the name given to the class of software that is surreptitiously installed on a computer, monitors user activity, and reports back to a third party on that activity.

Droppers: A dropper is a special kind of Trojan horse, the payload of which is to install a virus on the system under attack. The installation is performed on one or several infect able objects on the targeted system.

Injectors: An injector is a program very similar to a dropper, except that it installs a virus not on a program but in memory.

Germs: A germ is a program produced by assembling or compiling the original source code of a virus or of an infected program.

Directory Virus: A directory virus functions by infecting the directory of your computer. A directory is simply a larger file that contains information about other files and sub-directories within it. The general information consists of the file or directory name, the starting cluster, attributes, date and time and so forth. When a file is accessed, it scans the directory entry in search of the corresponding directory.

Hacker: Computer hacking refers to gaining unauthorized access to, and hence some measure of control over, a computer facility, and most countries now have specific legislation in place to deter those who might wish to practice this art and science. However, in practice, hackers generally have a particular target in mind, so their unauthorized access leads to further acts, which national law might also define as criminal activities. These can be summarized under the headings of unauthorized: Obtaining of confidential Information, Alteration or deletion of data and Code, Degradation or cessation of Services, Use of computer resources. [13]

Organization need to provide security skeleton to prevent the data hiding due to malicious code. In order to provide the security the organizations go through the security audit and most of the organization chose the internet security software as well as design their personal firewall and antivirus.

"**Antivirus**" is protective software designed to defend your computer against malicious software. Malicious software or Malware includes: viruses, Trojans, keyloggers, hijackers, diallers, and other code that vandalizes or steals your computer contents. Anyone who does a lot of downloading, or accesses diskettes from the outside world

on a regular basis should develop an antivirus strategy.

Antivirus software is equipped with features that not only check your files in your system, but also check your incoming and out-going e-mail attachments for viruses and other malicious programs [11]. The most important weapon in your antivirus is a clean, write-protected bootable system diskette. Booting from a clean write-protected diskette is the only way to start up your system without any viruses in memory. An effective defense against viruses is a clean backup of your hard drive. Many antivirus packages will attempt to disinfect infected programs for you so that the virus is no longer

in your system.

Antivirus products are categorized into three parts such as Internet Security [IS], Total Security [TS], and Antivirus [AV].

Antivirus products are the products, which are primarily focused on detecting and remediation viruses and Spyware.

Internet Security product provides all the virus and Spyware removal features of an AV, as well as additional functions to provide greater Internet protection. These features may include protection against phishing, root kit detection, firewalls and scanning of web pages and HTTP data.

Total Security: products provide data migration and backup features on top of all security features common to IS products. [5]

3 Performance Measurements of Security Tools

In order to measure the performance of IS, AV, and TS products we have taken five products for each category and few parameters for each category which are shown as below.

- **IS Product:**
  - Norton Internet Security 2011 [NIS] - Symantec Corporation.
  - Kaspersky Internet Security 2011 [KIS] - Kaspersky Lab.
  - AVG Internet Security 12.0-[AIS] AVG Technologies.
  - McAfee Internet Security 2011- [MIS] McAfee Inc.
  - Quick Heal Internet Security 2011- [QIS] Quick Heal Technologies.
- **TS Products:**
  - Norton 360 V5.0- [N360] Symantec Corporation.
  - Kaspersky Total Security 2011- [KTS] Kaspersky Lab.
  - Quick Heal Total Security 2011- [QTS] Quick Heal Technologies.
- **AV Products:**
  - Norton Antivirus 2011-[NA] Symantec Corporation.
  - Kapersky Antivirus 2011-[KA] Kaspersky Lab.
  - AVG Antivirus 12.0-AVG [AA] Technologies.
  - McAfee Antivirus 2011- [MA] McAfee Inc.
  - Quick Heal Antivirus 2011- [QA] Quick Heal Technologies.

Test Environment:

Above three category product are tested on a computer with the configuration as:

Pentium P-IV Core2Duo Processor, 1 GB RAM, Windows XP SP2 Operating System with applications such as MS-OFFICE 2007, MS-visual Studio 6.0,Turbo

C, Oracle 8 etc. An image of above OS and applications created with ghost software and fresh copy is installed to test the each category security tool. For the test, parameters used are:

1. INstallation Size. [INS] measured in MB
2. INstallation Time. [INT] measured in Sec
3. Boot Time. [BT] measured in Sec
4. Full Scan Time. [FST] measured in Minutes
5. User Interface Launch Time. [UILT] measured in Sec
6. Memory Utilization. [MU] measured in MB

Installation Size: The total installed size of the product

Installation Time: The time required installing the tool.

Boot time: The time taken for the machine to boot. Shorter boot times indicate that the application has less impact on the normal operation of the machine.

Scan Speed: The amount of time required to scan a typical set of clean files. 593 MB memory used for scanning.

User Interface Launch Speed: The time taken to start the User Interface of the product was measured.

Memory Utilization: The amount of RAM used by the product was measured while the machine and product were in an idle state, running in the background. All processes used by the application were identified and the total RAM usage calculated.

The performance of the product measured separately using each parameter and finally overall performance has been measured using score sheet the score points for each parameters are shown in table 1.

Table 1: Score point of the parameter

Sr. No.	Parameter	Score

1	Installation Size. [INS]	10
2	Installation Time. [INT]	10
3	Boot Time. [BT]	10
4	Full Scan Time. [FST]	10
5	User Interface Launch Time. [UILT]	10
6	Memory Utilization. [MU]	10

Table 2: Parameter wise test result of Internet Security Product

Parameter↓	**Product→**	**NIS**	**KIS**	**AIS**	**MIS**	**QIS**	**Unit**
INS		64.9	35.6	52.1	19.66	140	**MB**
INT		125	229	210	250	97	**Second**
BT		80	109	190	105	85	**Second**
FST		247	2759	1008	317	213	**Minute**
UILT		12	14	3	17	7	**Second**
MU		16.4	24	21.6	36.4	98.6	**MB**

Table 3: Parameter wise test result of Total Security Product

Parameter↓	**Product→**	**N360**	**MTS**	**QTS**	**Unit**
INS		167	105	382	**MB**
INT		162	387	440	**Second**
BT		133	121	214	**Second**

FST	318	359	337	**Minute**
UILT	2	11	5	**Second**
MU	14.5	66.4	69.3	**MB**

Table 4: Parameter wise test result of Antivirus Product

Parameter↓ Product→	**NA**	**KA**	**AA**	**MA**	**QA**	**Unit**
INS	32.8	31.7	50.5	218	325	**MB**
INT	85	195	100	260	370	**Second**
BT	131	110	85	200	116	**Second**
FST	612	266	319	256	320	**Minute**
UILT	5	5	6	2	4	**Second**
MU	31	19.8	16.5	35.9	30.6	**MB**

The Score point [SP] of each parameter for Internet Security evaluated as below:

INS SP = (IS product) * 10 / maximum of INS

INT SP = IS product) * 10 / maximum of INT

BT SP = IS product) * 10 / maximum of BT

FST SP = IS product) * 10 / maximum of FST

UILT SP = IS product) * 10 / maximum of UILT

MU SP = IS product) * 10 / maximum of MU

The Score point [SP] of each parameter for Total Security evaluated as below:

INS SP = (TS product) * 10 / maximum of INS

INT SP = (TS product) * 10 / maximum of INT

BT SP = (TS product) * 10 / maximum of BT

FST SP = (TS product) * 10 / maximum of FST

UILT SP = (TS product) * 10 / maximum of UILT

MU SP = (TS product) * 10 / maximum of MU

The Score point [SP] of each parameter for Antivirus Product evaluated as below:

INS SP = (AV product) * 10 / maximum of INS

INT SP = (AV product) * 10 / maximum of INT

BT SP = (AV product) * 10 / maximum of BT

FST SP = (AV product) * 10 / maximum of FST

UILT SP = (AV product) * 10 / maximum of UILT

MU SP = (AV product) * 10 / maximum of MU

Table 5: The conversion of tested figures into score point of the Internet Security Product

Parameter↓ Product→	NIS	KIS	AIS	MIS	QIS
INS	5	3	4	1	10
INT	5	9	8	10	4
BT	4	6	10	6	4
FST	1	10	4	1	1

UILT	7	8	2	10	4
MU	2	2	2	4	10

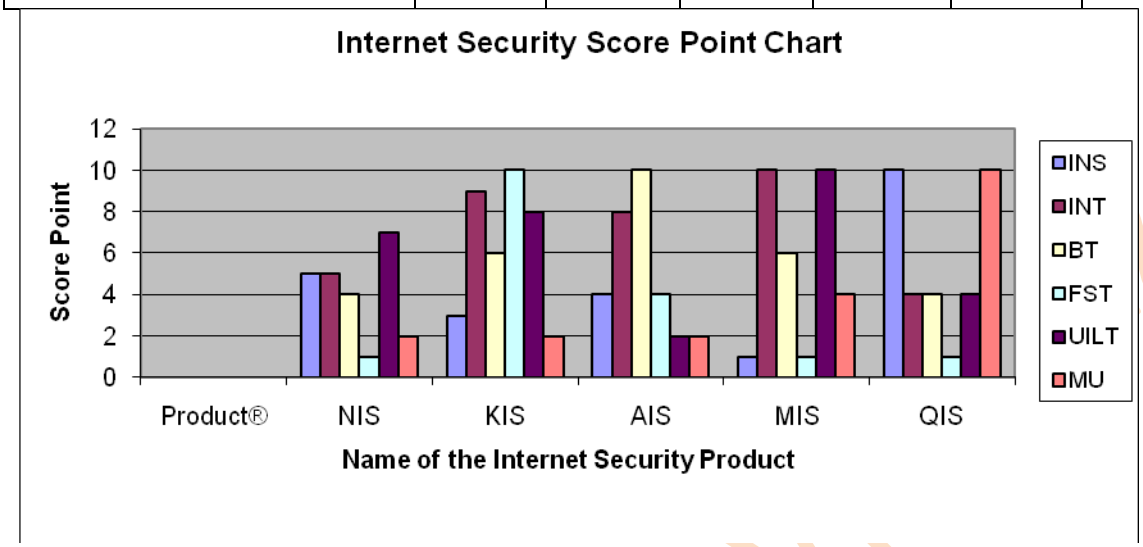


Chart 1: Parameter wise Rating of the Internet Security Product

Table 6: The conversion of tested figures into score point of the Total Security Product

Parameter↓ **Product→**	**N360**	**MTS**	**QTS**
INS	4	3	10
INT	4	9	10
BT	6	6	10
FST	9	10	9
UILT	2	10	5
MU	2	10	10

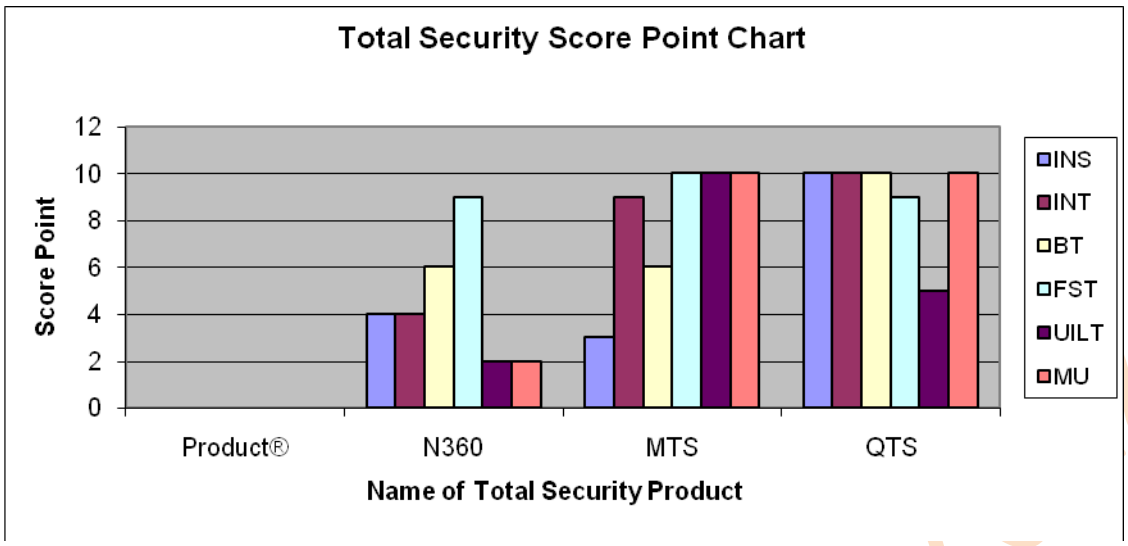


Chart 2: Parameter wise Rating of the Total Security Product

Table 7: The conversion of tested figures into score point of the Anti-Virus Product

Parameter↓ Product→	NA	KA	AA	MA	QA
INS	1	1	2	7	10
INT	2	5	3	7	10
BT	7	6	4	10	6
FST	10	4	5	4	5
UILT	8	8	10	3	7
MU	9	6	5	10	9

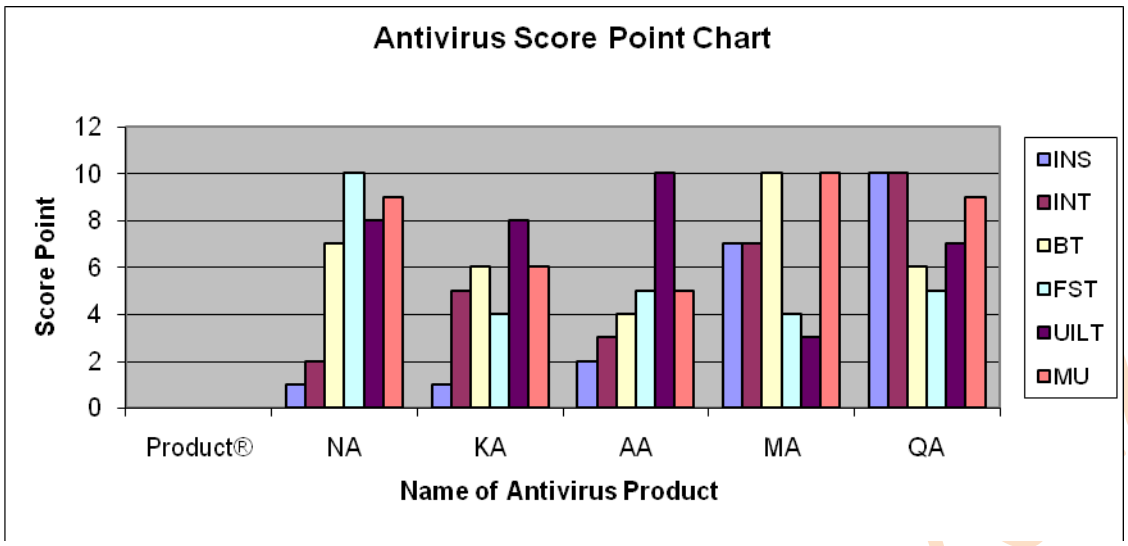


Chart 3: Parameter wise Rating of the Anti-Virus Product

The Total score point of each category is calculated as below

Total Score [TS] =sum of score point of each parameter

$$TSp = \sum_{n=1}^{6} Sp$$

Therefore score point of Internet Security of each product is as below

$$TS_{IS} = \sum_{n=1}^{6} S_{IS}$$

$$TS_{IS} = S_{INS}+S_{INT}+S_{BT}+S_{FST}+S_{UILT}+S_{MU}$$

Hence the Score Point of Norton Internet Security is,

Norton Internet Security TS_{IS}

$$= S_{INS}+S_{INT}+S_{BT}+S_{FST}+S_{UILT}+S_{MU}$$

= 5+5+4+1+7+2

= 24

Similarly, Total Score of each product is calculated. Accordingly the following table shows the total score point of each product.

In order to calculate the rating Score Point of the product has subtracted from 100 as,

Product Performance = 100 - Total Score Point

Therefore,

Norton Internet Security Performance = 100 - 24= 76

From the above calculation the total score point and Performance of each category is as below,

Table 8: The Total Score Point and Performance of the Internet Security Product

Product	NIS	KIS	AIS	MIS	QIS
Total Score	24	38	30	32	33
Performance	76	62	70	68	67

Table 9: The Total Score Point and Performance of the Total Security Product

Product	N360	MTS	QTS
Total Score	27	48	54
Performance	73	52	46

Table 10: The Total Score Point and Performance of the Anti-Virus Product

Product	NA	KA	AA	MA	QA
Total Score	37	30	29	41	47
Performance	63	70	71	59	53

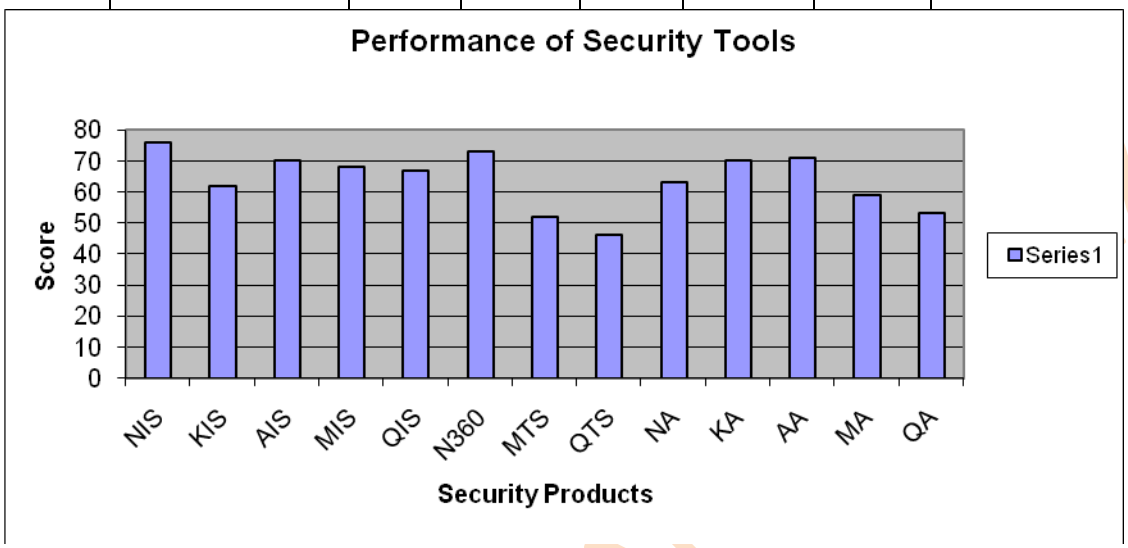


Chart 4: Comparison of Performance of all security tools

4 CONCLUSIONS

A comparative study of selected security tools was conducted on the basis of six parameters in order to explore their effectiveness and efficiency. The security tools selected were divided into three categories such as Total Security, Internet Security and Anti-virus Product. Each security tool was installed on a computer with Windows XP SP2 operating system, and then tested with the parameters and observations have been registered. The observations are with

different units and they need to be converted into Score Point, a unique scale. After that a comparison of performance of each tool is studied using likert-scale method. From the findings of the study, it is observed that four out of thirteen tools belongs to Good Category and remaining nine security tools falls within Very good Category, which has been shown in

following table.

Performance Of Security Tools				
Mediocre	**Fairly Good**	**Good**	**Very Good**	**Excellent**
0-20	**21-40**	**41-60**	**61-80**	**81-100**
-----	-----	MTS	NIS	-----
			KIS	
			AIS	
		QTS	MIS	
			QIS	
			N360	
		MA	NA KA	
		QA	AA	

From the above Chat 4 shows that NIS, N360, NA are performs well against all selected security tools.

5 ACKNOWLEDGMENTS

The researchers are grateful to the authors, writers, and editors of the books and articles, which have been referred for preparing the presented research paper. Reserchers are also greatful to Research Guide Dr. Prasanna Deshmukh and

Prof. Manikrao L. Dhore for their valuable guidance. It is the duty of researcher to remember their parents whose blessings are always with them.

6 REFERENCES

[1]. The art of computer virus research and defence by Peter Szor, Addison Wesley Professional, ISBN: 0-321-30454-3

[2]. Dr. Solomon's Virus Encyclopedia, 1995, ISBN 1897661002

[3]. Dr. Klaus Brunnstein 1999, from Antivirus to Antimalware Software and Beyond, http://csrc.nist.gov/nissc/1999/proceeding/papers

/p12.pdf

[4]. University of Tampere Dissertation 2002 By. M Helenius *acta.uta.fi/pdf/951-44-5394-8.pdf*

[5]. Paul Royal, Mitch Halpin, David Dagon, Robert Edmonds, and Wenke Lee. PolyUnpack: Automating the Hidden-Code Extraction of Unpack-

Executing Malware. In The 22th Annual Computer Security Applications Conference (ACSAC 2006), Miami Beach, FL, December

2006.

[6]. David Wren, Michael Fryer, Antivirus & Internet Security Performance Benchmarking 06/06/08

[7]. Rainer Link, Prof. Hannelore Frank, August, 2003, Server-based Virus-protection On Unix/Linux

[8]. Evgenios Konstantinou, Dr. Stephen Wolthusen, 2008, Metamorphic Virus: Analysis and Detection

[9]. Peter Szor, The Art of Computer Virus Research and Defense. Addison Wesley Professional, 1 edition, February 2005.

[10]. Thomas F. and Andrew Urbaczewski, Spyware: The ghostIn the machine. Communications of the Association for Information Systems,

14:291{306, 2004.

[11]. Felix Uribe, Protecting your Personal Computer against Hackers and Malicious Codes

[12]. Bob Kanish, by "An Overview of Computer Viruses and Antivirus Software"

[13] N. Nagarajan, The basics of protecting against computer hacking

IT 034

Application of cloud computing on E-learning

Sachin V. Joshi **Shrutika V Hazare** **Devendra R. Bandbuche**
Dept. of MCA **Dept. of MCA** **Dept. of MCA**
P.R.M.I.T.&R,Badnera **P.R.M.I.T.&R,Badnera** **P.R.M.I.T.&R,Badnera**

Abstract:

Cloud computing builds on decades of research in virtualization, distributed computing, utility computing, and more recently networking, web and software services. E-learning systems usually require many hardware and software resources. The biggest players in the field of e-learning software have now versions of the base applications that are cloud oriented. However, the current models of e-learning ecosystems lack the support of underlying infrastructures, which can dynamically allocate the required computation and storage resources for e-learning ecosystems. Cloud computing infrastructure and related mechanisms allow for the stability, equilibrium, efficient resource use, and sustainability of an e-learning ecosystem. Hence, this paper introduces Cloud computing into an e-learning ecosystem as its infrastructure. In this paper, an e-learning ecosystem based on Cloud computing infrastructure is presented.

Keywords
Distributed computing, utility computing, grid computing, Reliability, Scalability.

1 .Introduction

Cloud Computing has been one of the most booming technology among the professional of Information Technology and also the Business due to its Elasticity in the space occupation and also the better support for the software and the Infrastructure it attracts more technology specialist towards it. Cloud plays the vital role in the Smart Economy, and the possible regulatory changes required in implementing better Applications by using the potential of Cloud Computing .The main advantage of the cloud is that it gives the low cost implementation for infrastructure and some higher business units like Google, IBM, and Microsoft offer the cloud for Free of cost for the Education system, so it can be used in right way which will provide high quality education [1][2][3]. Cloud computing is Internet-("cloud") based development and use of computer technology ("computing").The idea of cloud computing is based on a very fundamental principle of “re usability of IT capabilities”.

According to IEEE computer society cloud computing is a paradigm shift whereby details are abstracted from the users who no longer have need of, expertise in, or control over the technology infrastructure "in the cloud" that supports them. Cloud computing describes a new supplement, consumption and delivery model for IT services based on the Internet, and it typically involves the

provision of dynamically scalable and often virtualized resources as a service over the Internet. A technical definition is "a computing capability that provides an abstraction between the computing resource and its underlying technical architecture (e.g., servers, storage, networks), enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort. Many education institutions do not have the resources and infrastructure needed to run top e-learning solution. This is why Blackboard and Moodle, the biggest players in the field of e-learning software, have now versions of the base applications that are cloud oriented.

2. Pretended scenario

In this scenario cloud computing is being looked upon by experts in various domains because of its advantages. Cloud has been used in the business oriented unit and in the current education system in India the teaching via web is not so widely available and adapted. Even if it is available, it is provided at a very high cost. Cloud has generated many resources which can be used by various educational institutions and streams where their existing/proposed web based learning systems can be implemented at low cost.

A. Benefits of Cloud Computing

The advantages that come with cloud computing can help resolving some of the common challenges one might have while supporting an educational institution [6].

➢ **Cost**

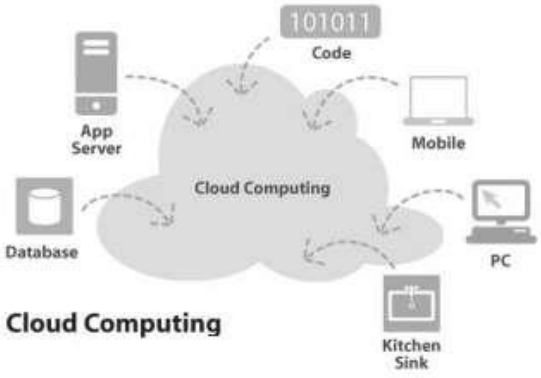
One can choose a subscription or in, some cases, pay-as-you go plan – whichever works best with that organization business model.

➢ **Flexibility**

Infrastructure can be scaled to maximize investments. Cloud computing allows dynamic scalability as demands fluctuate.

➢ **Accessibility**

This help makes data and services publicly available without make vulnerable sensitive information.



➢ **The Client – The End User**

Everything ends with the client (mobile). The hardware components, the application and everything else developed for

Cloud computing will be used in the client. Without the client, nothing will be possible. The client could come in two forms:

the hardware component or the combination of software.

B. Services in cloud computing

Infrastructure as a Service**.** One can get on-demand computing and storage to host, scale, and manage applications and services. Using Microsoft data centers. In the previous generation of the information technology the data sharing which led the path for the knowledge sharing was not used by the users globally, in this generation the various streams have the knowledge of e-Learning and the Mobile based learning. In this present context the usage of the central data centre is a easy process for the education system however the cost of implementation and the maintenance of the data storage space and also the load capability also software licensing depends on the real time usage of these systems. Business streams can make revenue out of those expenses whereas for educational institutions which really want to motivate the learners and want to offer a quality education at affordable cost can achieve

this by spending a large amount. This can be overcome by the present cloud computing technology that is "Pay as Use" (PAU).

3. E-learning based cloud computing

Internet has had a profound impact on the way we interact and work. It started of as a medium to exchange information between computers. With increasing network speeds and wider penetration of Internet, software providers started moving their applications to the Web, and started offering their software's as a service (SaaS). Organizations or individuals could now use enterprise level software's for a small fee without having to install them on local infrastructure, and without worrying about how to maintain these applications.But all this while there was almost no innovation on how hardware infrastructure was deployed or licensed. One either had to create an in-house data center to deploy and manage their software systems, or had to take servers on rent in a data-center managed by someone.

4. E-learning benefits

E-learning is widely used today on different educational levels: continuous education, company trainings, academic courses, etc.

There are various e-learning solutions from open source to commercial. There are at least two entities involved in an e-learning system: the students and the trainers[5].

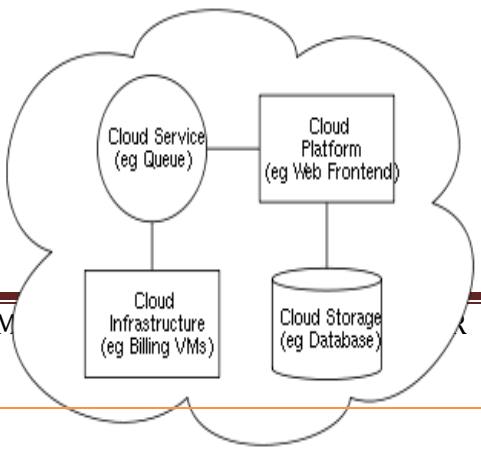
E-learning systems can use benefit from cloud computing using:

- **Infrastructure**: use an e-learning solution on the provider's infrastructure
- **Platform**: use and develop an e-learning solution based on the provider's development interface
- **Services**: use the e-learning solution given by the provider.

A very big concern is related to the data security because both the software and the data are located on remote servers that can crash or disappear without any additional warnings.

Even if it seems not very reasonable, the cloud computing provides some major security benefits for individuals and companies that are using/developing e-learning solutions, like the following:

- **improved improbability** – it is almost impossible for any interested person (thief) to determine where is located the machine that stores some wanted data (tests, exam questions, results) or to find out which is the physical component he needs to steal in order to get a digital asset;



Cloud Computing Sample Architecture

- **Virtualization** – makes possible the rapid replacement of a compromised cloud located server without major costs or damages[3]. It is very easy to create a clone of a virtual machine so the cloud downtime is expected to be reduced substantially;
- **centralized data storage** – losing a cloud client is no longer a major incident while the main part of the applications and data is stored into the cloud so a new client can be connected very fast. Imagine what is happening today if a laptop that stores the examination questions is stolen[5].

5. Cloud Computing Architecture

Cloud architecture, the system architecture of software system involve in the delivery of cloud computing comprises hardware and software designed by the cloud architect who typically works for cloud integrator. It typically involves multiple cloud components communicating with each other over application programming interface, usually web services. This closely resembles the Unix philosophy of having multiple programs each doing one thing well and working together over universal interfaces. Complexity is controlled and the resulting systems are more manageable than their monolithic counterparts. Cloud architecture extends to client, where web browser and software application access cloud applications. storage architecture is loosely coupled, often assiduously avoiding the use of centralized metadata servers which can become bottlenecks. This enables the data nodes to scale into the hundreds, each independently delivering data to applications or users.

6.Characteristics

1) In general, cloud computing customers do not own the physical infrastructure, instead avoiding capital expenditure by renting usage from the third party provider.

2) They consume resources as a service and pay only for the resources that they use.

3) Sharing "perishable and intangible" computing power among multiple tenants can prove utilization rates, as servers are not unnecessarily left idle.

4) A side effect of this approach is that overall computer usage rises dramatically, as customers do not have to engineer for peak load limits.

7. Cloud computing types

Public Cloud

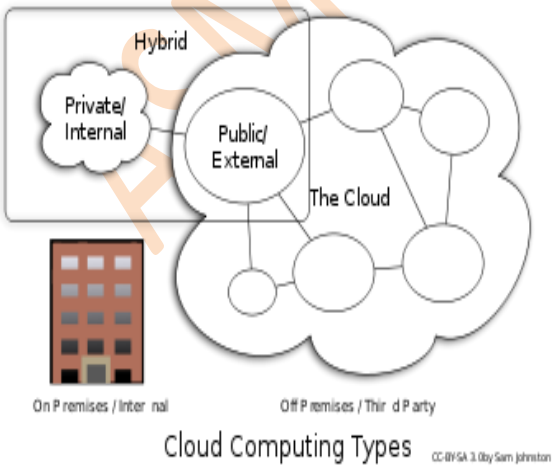
Public cloud or external cloud describes cloud computing in the traditional mainstream sense, whereby resources are dynamically provisioned on fine grained, self services basis over the internet via web applications, web services from an offsite third party provider who shares resources and bills on a fine-grained utility computing basis.

Hybrid cloud

A hybrid cloud environment consisting of multiple internal and/or external providers" will be typical for most enterprises". A hybrid cloud can describe configuration combining a local device, such as a Plug computer with cloud services. It can also describe configurations combining virtual and physical, collocated assets—for example, a mostly virtualized environment that requires physical servers, routers, or other hardware such as a network appliance acting as a firewall or spam filter.

Private cloud

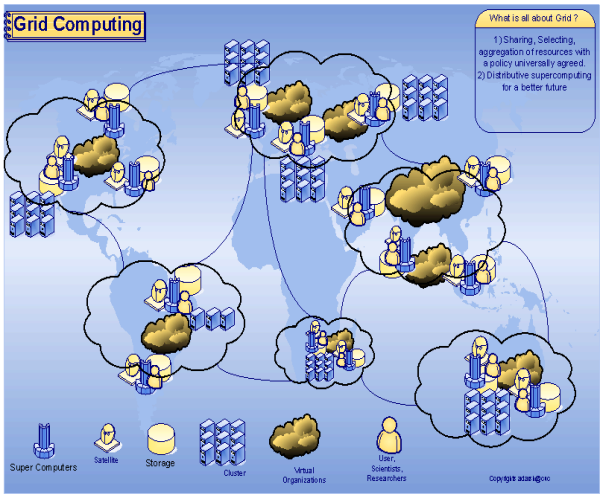
Private cloud and internal cloud are neologisms that some vendors have recently used to describe offerings that emulate cloud computing on private networks. These products claim to "deliver some benefits of cloud computing without the pitfalls", capitalizing on data security, corporate governance, and reliability concerns. They have been criticized on the basis that users "still have to buy, build, and manage them" and as such do not benefit from lower up-front capital costs and less hands-on management, essentially" the economic model that makes cloud computing such an intriguing concept".



8. Grid computing

Grid computing is a term for either of two broad subcategories of distributed computing:
1 Online computation or storage offered as a service supported by a pool of distributed computing resources, also known as utility computing, on-demand computing, or cloud computing. Data grids provide controlled sharing and management of large amounts of distributed data, often used in combination with computational grids. **2** The creation of a "virtual supercomputer"

composed of a network of loosely-coupled computers, acting in concert to perform very large tasks. This technology has been applied to computationally-intensive scientific, mathematical, and academic problems through volunteer computing, and it is used in commercial enterprises for such diverse applications as drug discovery, economic forecasting, seismic analysis, and back-office data processing in support of e-commerce and web services



9.Difference between Cloud computing & grid computing

- Grid computing emphasizes on resource sharing, every grid node can apply for resource from other nodes, and every node should contribute resource to the grid. The focus of grid computing is on the ability of moving a workload to the location of the needed computing resources, which are mostly remote and are readily available for use. Grids also require applications to conform to the grid software interfaces.
- Cloud computing emphasize on proprietary, every user out of the cloud can get it's own private resource from the cloud, and the cloud resource are provided by the specific service provider, the user need not contribute its resource. In a cloud environment, computing resources, such as servers, can be dynamically shaped or carved out from its underlying hardware infrastructure and made available to a workload. In addition, while a cloud does support grid, a cloud can also support nongrid environments, such as a three-tier Web architecture running traditional or Web 2.0 applications.
- Grid computing emphasizes on computing sensitive task, and is difficult to automated scale. Cloud computing emphasizes on transactional application, a great amount of separate request, and can scale automatically or semi automatically.

10.POTENTIAL ISSUES

- Privileged user access.
- Regulatory compliance.
- Data location.
- Data segregation.
- Investigative support.
- Long-term viability.

11.Conclusions

The development of e-learning solution cannot ignore the cloud computing trends.

There are many benefits from using the cloud computing for e-learning systems. Using cloud computing for e-learning solutions influences the way the e-learning

Software projects are managed. There are specific tasks that deal with finding providers for cloud computing, depending on the requirements (infrastructure, platform or services). Also, the cost and risk management influences the way the e-learning solutions based on cloud computing are managed. The extent of some public clouds across multiple legal jurisdictions further complicates this issue; see "Legal Issues" for more detail. These concerns are considered key obstacles to broader adoption of cloud computing, making them areas of active research and debate among cloud computing practitioners and advocate. The cloud computing has the significant scope to change he whole education system. In present scenario the e-learning is getting the popularity and this application in cloud computing will surely help in the development of the education offered to poor people which will increase the quality of education offered to them. Cloud based education will help the students, staff, Trainers, Institutions and also the learners to a very high extent and mainly students from rural parts of the world will get an opportunity to get the

knowledge shared by the professor on other part of the world. Even governments can take initiatives to implement this system in schools and colleges in future and we believe that this will happen soon[4].

References

[1] Bacigalupo, David; Wills, Gary; De Roure, David; *Victor, A Categorization of Cloud Computing Business Models: IEEE/ACM May 2010.*

[2] Minutoli, G. Fazio, M. Paone, M. Puliafito, A. Engineering Fac, Univ.of Messina, Messina, Italy *Virtual Business Networks With Cloud Computing and Virtual Machines: IEEE/ICUMT Oct 2010.*

[3] Paul Hofmann, *SAP Labs,* Dan Woods, *CITO Research:* The Limits of Public Clouds for Business Applications: *Digital Library* November/December 2010.

[4] DeCoufle B.The impact of cloud computing in schools, The Datacenter Journal,http://datacenterjournal.com/content/view/3032/40/, July 2009

[5] Pocatilu P., Boja C.Quality Characteristics and Metrics related to M-Learning Process, Amfiteatru Economic, Year XI, June 2009

IT 035

Secure Mobile Ad hoc Network

Varsha S. Upare

Department of Computer Engg.

MIT College of Engg., Pune

varsha.upare@gmail.com

***Astract*-Mobile ad hoc network is the collection wireless mobile nodes that forms short term network. Security is an important in mobile ad hoc network.Manet is prone to security attack due to lack of centralized approach, dynamic network topology. In this paper, various attacks in the mobile ad hoc network have been studied. The malicious behaviour of a node and the need for the security in the ad hoc is defined in this paper. SCAN technology protects the network from data routing and data forwarding operrations.It guard the network from detecting and reacting to the malicious node. In SCAN each neighbouring node monitors each other, sustain each other, no node is better than other. The analysis of SCAN design has been studied in this paper.The survey analysis of packet delivery ratio,miss detection ratio with respect to mobilitiy has been covered.**

KEYWORD-Mobile ad hoc network, Nodes, Security, Attack, Malicious Behaviour.

I. INTRODUCTION

A MANET is a self-organizing system of mobile nodes that communicate with each other via wireless links with no fixed infrastructure or centralized administration such as base stations or access points. MANETs are suitable for applications in which no infrastructure exists such as military battlefield, emergency rescue, vehicular communications and mining operations. In these applications, communication and collaboration among a given group of nodes are necessary. In this paper, study of the effects of different types of attacks on mesh-based multicast in MANETs is considered. Here, the most common types of attacks, namely rushing attack, black hole attack, neighbour attack and jellyfish attack is considered. Ad Hoc Network provides quick communication among nodes to transfer the packets from one node to other. All the links between nodes are wireless. Any malicious node in the network can disturb the whole process. Whenever a node exhibits a malicious

behaviour under any attack, it assures the breach of security principles like availability, integrity, confidentiality etc [5]. An intruder takes advantage of the vulnerabilities ,presents in the ad hoc network and attacks the node which breaches the In this paper, we tackle an important security issue in ad hoc networks, namely the protection of their network-layer operations from malicious attacks. We focus on securing the packet delivery functionality. Several recent studies [1]–[4] have provided detailed description on such network- layer security threats and their consequences. In SCAN, each node monitors the routing and packet forwarding behaviour of its neighbours, and independently detects any malicious nodes in its own neighbourhood. The monitoring mechanism takes advantage of the broadcast nature of wireless

Communication. In a network with reasonable node density, one node can often overhear the packets (including both routing updates and data packets)

Received, as well as the packets sent by a neighbouring node. In such cases, it can *cross-check* these packets to discover whether this neighbour behaves normally in advertising routing updates and forwarding data packets.

This paper is organized as follows: Section II covers the Various attacks in the mobile ad hoc network. Section III covers the vulnerabilities present in the ad hoc network. Due to this vulnerability, node behaves in malicious manner. Section IV defines malicious behaviour. . Section V describes the SCAN design in details. Section VI analyzes the miss detection ratio. False accusation and packet delivery ratio with respect to mobility. Section VII concludes the paper.

II. ATTACKS

Rushing attack.. When source nodes flood the network with route discovery packets in order to find routes to the destinations, each intermediate node processes only the first non-duplicate packet and discards any duplicate packets that arrive at a later time. Rushing attackers, by skipping some of the routing processes, can quickly forward these packets and be able to gain access to the forwarding group. This type of attacks was first introduced in [4].

- **Black hole attack**. In the balckhole attack, the attacker simply drops all of data packets it receives. This type of attack often results in very low packet delivery ratio. Simply forwards the packet without recording its ID in the packet, it makes two nodes that are not within th communication range of each other believe that they are neighbours, resulting in a disrupted route.

- **Jellyfish attack**. Similar to the black hole attack, a jellyfish attacker first needs to intrude into the forwarding group and then it delays data packets unnecessarily for some amount of time before forwarding them. This result in significantly high end-to-end delay and delay jitter, and thus degrades the performance of real-time applications. Jellyfish attacks were first discussed by Aad *et al.* in [1].

• **Neighbouring attack**. Upon receiving a packet, an intermediate node records its ID in the packet before forwarding the packet to the next node. However, if an attacker simply forwards the packet without recording its ID in the packet, it makes two nodes that are not within the communication range of each other believe that they are neighbours resulting in a disrupted route.

III. NEED OF SECURITY IN AD HOC NETWORK

The security is important in mobile ad hoc network. Following are the some of the issues in the mobile ad hoc network.

Mobility- Each node in ad hoc network is movable. It can join or leave a network at any instant of time without informing any node. This gives chance to intruder to easily enter in the network or leave the network

Open Wireless Medium- All the communication among nodes is taking place through the medium of air an intruder can easily access medium to get

Information about the communication or can easily trap it.

Resource Constraint- Every node in mobile ad hoc network has limited resources like battery, computational power, bandwidth etc. An intruder can unnecessarily waste these limited resources in order to make it unavailable to perform.

Dynamic Network Topology- As the nodes are travel ,, the topology changes every time . The packets from source to destination may take different path for communication. An intruder can introduce itself in any path.

Scalability- Ad hoc network may contain of number of nodes. This number is not fixed. In a network of its range, as many as number of nodes can take part. Intruder simply takes advantage of this parameter as there is no limitation on number of nodes.

Reliability- All the wireless communication is limited to a range of 100 meter which puts a constraint on nodes to be in range for establishing communication. Due to this limited range, some data errors are also generated. For attacking a particular node, an intruder needs to be in its

range.

IV. MALICIOUS BEHAVIOUR OF A NODE

Malicious Behaviour- "When a node breaks the security principles and is under any attack. Such nodes show one or more of the following behaviour:

Packet Drop- Simply consumes or drops the packet and does not forward it.

Battery Drained- A malicious node can waste the battery by performing unnecessarily operations.

Buffer Overflow- A node under attack can fill the buffer with forged updates so that real updates cannot be stored further.

Bandwidth Consumption- Whenever a malicious node consumes the bandwidth so that no other legitimate node can use it.

Malicious Node Entering- A malicious node can enter in the network without authentication.

Stale Packets- This means to inject stale packets into the network to create confusion in the network.

Link Break- malicious node restricts the two legitimate communicating nodes from communicating.

Message Tampering- A malicious node can alter the content of the packets.

Fake Routing- Whether there exists a path between nodes or not, a malicious node can send forged routes to the legitimate nodes in order to get the packets or to disturb the operations.

Stealing Information- Information like the content, location, sequence number can be stolen by the malicious node to use it further for attack.

Session Capturing- When two legitimate nodes communicate, a malicious node can capture their session so as to take some meaningful information.

V. SCAN DESIGN

In this section, the SCAN design has been studied in details. To secure the packet delivery functionality, each SCAN Node overhears the wireless channel in the promiscuous mode, and observes the routing and packet forwarding behaviour of its Neighbours at all time. A malicious node is convicted when its neighbours have reached such a consensus, then it is removed from the network membership and isolated in the network. To be a part of the network, each legitimate node carries a valid token. Without a valid token node cannot participate in the network. A legitimate node can always renew the token from its neighbours before its current Token expires. However, when a malicious node is convicted, its neighbours collectively revoke its current token and inform all other nodes in the network. Scan frame work goes through the following process.

- *Collaborative monitoring*: All nodes within a local neighbourhood collaboratively monitor each other.

- *Token renewal*: All legitimate nodes in a local neighbourhood collaboratively renew the tokens for each other.

- *Token revocation*: The neighbours of a malicious node upon consensus collaboratively revoke its current token. In this framework, the malicious nodes are detected and convicted via the collaborative monitoring mechanism.

A. Token Renewal

The token renewal mechanism Guarantees that legitimate nodes can persist in the Network by renewing their token from time to time. To participate in the network, each legitimate node carries a token which contains the following three fields (owner_time, signing_time, expiration_time). The tokens are protected by the public-key cryptographic mechanism. Before the current token expires, each node request its local neighbours to renew its token. The node that needs token renewal broadcasts a token request (TREQ) packet, which contains its current token and a timestamp. each node keeps a token revocation list (TRL) based on the token revocation mechanism. Specifically, when a node receives a TREQ packet from its neighbour, it removes the token from the packet. It checks whether the token has already been revoked by comparing it with the TRL. If the token is still valid yet about to expire, it constructs a new token with owner_id equal to that in the old token signing_time equal to the timestamp in the TREQ packet. the expiration_time is determined by the credit strategy defined as below.

1) Credit Strategy in Token Lifetime:

Here the credit strategy of a token is explained. In this strategy, a newly joined node is issued a token with short lifetime. It collect Its credit when it remains to behave well in the network and its subsequent token lifetime depends on its credit at the renewal time. The more credit one node has, the longer lifetime its token has. This way, a legitimate node will have its token lifetime steadily increased over time, thus renewing its token less and less frequently.

B. Collaborative Monitoring

The collaborative monitoring mechanism in SCAN observes the routing and packet forwarding operations of each node in a fully decentralized and localized manner. Each node overhears the channel, monitors the behaviour of its neighbours, and discovers consistent misbehaviour as indications of attacks. Moreover, local neighbouring nodes collaborate with each other to improve the monitoring accuracy.

1) Monitor Routing Behaviour:

Our basic idea is to overhear the channel and *cross-check* the routing messages announced by different nodes. This can be applied to any distributed and deterministic routing protocol. In such protocols, the routing activity of a node is a three-step process: 1) receiving routing updates from neighbouring nodes as inputs to a routing algorithm

2) Executing the routing algorithm; and 3) announcing the output of the routing algorithm as its own routing updates. The monitoring task is to verify whether the routing algorithm executed by a node follows the protocol specifications. In other words, the trustworthiness of a routing message, as the output of the routing algorithm, can be examined when the monitoring node knows the input to the algorithm, since the algorithm itself is publicly known and deterministic. This idea can be illustrated with the context of AODV. An AODV node cannot acquire enough information about the routing algorithms. The key reason is that the *next hop* information is missing in the AODV routing messages. Thus, when a node announces a routing update, its neighbours have no hint about which node is the next hop in the route and, hence, cannot judge on its input to the routing algorithm, i.e., the original routing update on which its routing computation is based .In order to permit the cross-checking of routing updates, two modifications to AODV need to be done. First, add one more field, *next_hop*, in the RREP packet. Similarly, add one more field, *previous_hop*, in the RREQ packet. This way, each node explicitly claims its next hop in a route when it advertises routing updates. Second, each node keeps track of the routing updates previously announced by its neighbours. Essentially, each node maintains part of the routing tables of its neighbours.

2) Monitor Packet Forwarding Behaviour:

Each SCAN node also monitors the packet forwarding activity of its neighbours. This is achieved by overhearing the channel and comparing ongoing data transmission with previously recorded routing messages. The currently focus has been on three kinds of forwarding misbehaviour, namely, packet drop, packet duplication, and network-layer packet jamming, and develop simple algorithms to detect each of them. Packet drop means that a node drops the packets that it is supposed to forward for its neighbours; packet duplication means that a node duplicates the packets that it has already forwarded; and packet jamming means that a node sends too many packets and occupies a significant portion of the channel bandwidth In ,The packet drop detection algorithm in which the sender explicitly lists the route in the data packet header. It cannot be directly applied in the AODV context, because when a node receives a packet, its neighbours do not know to which node it should forward the packet, thus, cannot tell whether it forward the packet in the correct manner. Fortunately, our modification to the AODV protocol, described in the previous section, enables the detection of packet drop, because each node keeps track of the route entries announced by its neighbours, which explicitly lists the *next_hop* field. Specifically, each SCAN node records the headers of the recent packets it has overheard. If one node overhears that the bandwidth consumed by duplicate packets from its neighbour exceeds the threshold *Duplicate_Bandwidth*, or the bandwidth consumed by packets originated from its neighbour exceeds the threshold *Sending Bandwidth*, it also considers these events as packet forwarding misbehaviour.

VI. SIMULATION EVALUATION

In this section, we evaluate the performance of SCAN through extensive simulations, the goal of which is to answer the following questions

In the simulations, the following metrics are observed: 1) *miss detection ratio*, which is the chance that SCAN fails to convict and isolate a malicious node; 2) *false accusation ratio*, which is the chance that SCAN incorrectly convicts and isolates a legitimate node; 3) *packet delivery ratio*, which is the percentage of packets that are successfully delivered to the receiver nodes; and 4) *communication overhead*, which is the total number of packets sent by SCAN in order to achieve its goal.

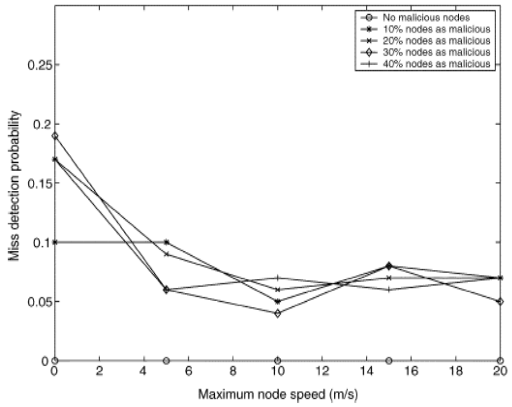


Fig. 1. Miss detection probability versus mobility.

the miss detection ratio obtained by considering only the set of active malicious nodes, instead of all prechosen malicious nodes. The false accusation ratio is obtained in a similar way over the set of active legitimate nodes

B. Monitoring and Detection

The detection performance of the collaborative monitoring mechanism in SCAN in terms of miss detection and false accusation ratios .Fig 1 shows the miss detection ratio as the node mobility speed changes. Ratio is the highest in a static network, regardless of the number of malicious nodes. The miss detection ratio drops considerably when nodes start to move, and remains stable at 4%–8% when the speed further increases. SCAN fails to convict a malicious node mainly because it resides in a sparsely occupied region. In a static network, if a malicious node happens to stay in a sparsely occupied region, its neighbours always have no chance to convict it. On the contrary, in a mobile network, the mobility increases the chance that other nodes roam into this region or the malicious node itself moves into another densely occupied region. The impact of node mobility on the false accusation ratio is presented in Fig. 7 the false accusation ratio continues to increase as nodes move faster. When the maximum speed is 20 m/s, the false accusation ratio is around 5% 10%. The reason is that higher mobility makes nodes more

"memory less. Fig. 6 and 7 also illustrate the impact of the number of malicious nodes on the detection performance. In both cases, even if the number of malicious nodes increases dramatically from 0% to 40% of the network population, it does not exhibit evident impact on the detection performance.

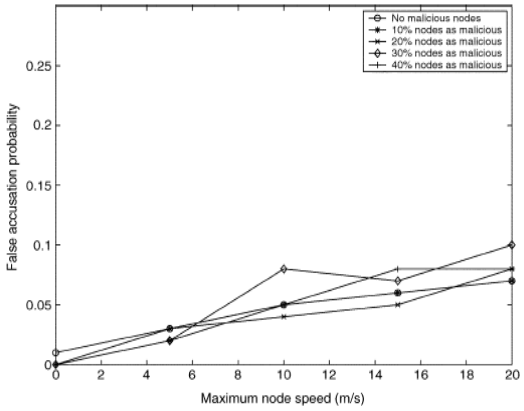


Fig. 2. False accusation probability versus mobility

C. Packet Delivery Ratio

Fig. 8 shows the improvement on the packet delivery ratio in a SCAN-protected network. In these simulations, 30% of the nodes are set as malicious nodes. From the figure that SCANS increases the packet delivery ratio by a factor up to 150% even if 30% of nodes are malicious. In an ad hoc network without any security protection, the packet delivery ratio can be as low as 30%, even if the network is lightly loaded. Another observation from Fig. 3 is that even in a SCAN protected and light-loaded network, the packet delivery ratio is not 100%.

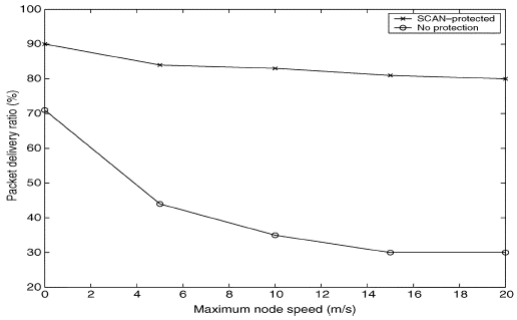


Fig. 3. Packet delivery ratio versus mobility.

VII. CONCLUSION

Here in this paper, the study of various attack has been done. Mobile ad hoc network is vulnerable to various attack Some of the vulnerabilities of the

Mobile ad hoc network has been studied. This paper explores a novel self-organized approach to securing such networks. To this end, SCAN technology is presented, a network-layer security solution that protects routing and forwarding operations in a unified framework. SCAN exploits localized collaboration to detect and react to security threats. All nodes in a local neighbourhood collaboratively monitor each other and sustain each other, and no single node is superior to the others. Both analysis and simulations results have confirmed the effectiveness and efficiency of SCAN in protecting the network layer in mobile ad hoc networks

REFRENCES

[1] S. Marti, T. Giuli, K. Lai, and M. Baker, "Mitigating routing Misbehaviour in mobile ad hoc networks," in *Proc. ACM MobiCom*, 2000, pp. 255–265.

[2] J. Hubaux, L. Buttyan, and S. Capkun, "The quest for Security in mobile ad hoc networks," in *Proc. ACM MobiHoc*, 2001, pp. 146–155.

[3] J. Kong, P. Zerfos, H. Luo, S. Lu, and L. Zhang, "Providing Robust and ubiquitous security support for MANET," in *Proc. IEEE ICNP*, 2001, pp. 251–260.

[4] Y. Hu, A. Perrig, and D. Johnson, "Ariadne: A secure on-Demand routing protocol for ad hoc networks," in *Proc. ACM MobiCom*, 2002, pp. 12–23.

[5] Y.C. Hu, A. Perrig, and D.B. Johnson, "Rushing Attacks And Defense in Wireless Ad Hoc Network Routing Protocols", Proceedings *of ACM WiSe 2003*, San Diego, CA, Sep. 2003

[6] M. Zapata and N. Asokan, "Securing ad hoc routing Protocols," in *Proc. ACM WiSe*, 2002, pp. 1–10.

[7] C. Perkins and E. Royer, "Ad hoc on-demand distance vector routing," in *Proc. IEEE WMCSA*, 1999, pp. 90–100.

[8] C. Perkins, E. Royer, and S. Das, "Ad hoc on demand Distance vector (AODV) Routing," Internet Draft, draft-ietf manet-aodv-10.txt, 2002.

[9] Y. Hu, A. Perrig, and D. Johnson, "Packet leashes: A Defense against wormhole attacks in wireless ad hoc networks," in *Proc. IEEE INFOCOM*, 2003, pp. 1976– 1986.

[10] L. Buttyan and J. Hubaux, "Stimulating cooperation in self-organizing mobile ad hoc networks," *ACM/Kluwer Mobile Netw. Applicat.*, vol. 8, no. 5, pp. 579–592, Oct. 2003.

IT 036

WEB SERVICE TESTING TOOL

Rohit Kishor Kapadne
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id- kapadnerohit@gmail.com

Contact no.7709400608

Ishwar M. Mali
College of Engineering, Manjari
Hadapsar, Dist.Pune
State-Maharashtra
Country-India
Email id- maliishwar05@gmail.com

Contact no.8600285228

A.Abstract

Web Services is arguably the most powerful and popular software technology in today's connected e-world. This application gives a briefing about web services and also touches upon the various standards that have evolved over the years, such as XML, SOAP, WSDL, and UDDI. This paper advocates the need for testing web services as Testing Web Services poses a big challenge to testing professionals because of its inherently complex and distributed nature. Web Service Testing Tool(WSTT) will enable us to locate and invoke web service methods directly. It supports all of the core Web service technologies like WDSL, SOAP, and it is an ideal Web service tool for testing Web services, inspecting WSDL files, automating or accelerating verification of each component when developing Web service enabled applications.

To call the web services which deployed on application/web Server, http protocol should be used. However there are web services developed using technologies like MQ [Message Queing] and WMB [Web sphere Message Broker] which deploys the web services on Broker and expects input in the Queue instead of http request.

There are testing tools available which either supports http exposed web services or MQ based web services. And also the existing testing tools are Net tool, RFHUTIL; SOAPUI supports only one request to be hit at a time to the web service.

B.Introduction-

Web Service is a SOA (Service Oriented Architecture) technology used to expose any business functionality as a service over the web.

To call the web services which deployed on application/web Server, http protocol should be used. However there are web services developed using technologies like MQ [Message Queing] and WMB [Web sphere Message Broker] which deploys the web services on Broker and expects input in the Queue instead of http request.

Innovativeness:

Make Web Services Testing Tools an integral part of your development life-cycle to ensure application robustness and interoperability without writing custom testing code.

Our developed application tool enables us to test the scalability and robustness of our Web Services. Test runs can be set for any WSDL operation by enabling Performance Mode in our application. All previously authored regression tests can be used for performance and scalability testing. We can run test for a specified duration to ensure that Web Services are robust.

We can also run the test for a specified number of iterations to measure Web Services performance and scalability. Up to 100 concurrent parallel request can be setup to measure scalability. Our developed tool provides additional test agents to scale beyond 100 concurrent request. Each tool user provides 100 additional request or test cases. There is no limit to the number of users that can be managed by our Testing Tool.

All security and identity functionality is also available for testing your Web Services using developed tool.

11111

Logic behind:

Web services are more and more used for the realization of distributed applications crossing domain borders. However, the more Web services are used for central and/or business critical applications, their functionality, performance and overall quality become key elements for their acceptance and wide spread use. Consumers of Web services will want assurances that a Web service will not fail to return a response in a certain time period. Even

more, systematic testing of Web services is essential as Web services can be very complex and hard to implement: although the syntax of the data formats is described formally with XML, the semantics and possible interactions with the Web service and use scenarios are described textually only. This encompasses the risk of misinterpretation and wrong implementation. Therefore, testing a final implementation within its target environment is essential to

assure the correctness and interoperability of a Web service.

Testing a system is performed in order to assess its quality and to find errors if existent. An error is considered to be a discrepancy between observed or measured values provided by the system under test and the specified or theoretically correct values. Testing is the process of exercising or evaluating a system or system component by manual or automated means to check that it satisfies specified requirements. Testing approves a quality level of a tested system. The need for testing approaches arose already within the IT community: so-called interoperability events are used to evaluate and launch certain XML interface technologies and Web services, to validate the specifications and to check various implementations for their functionality and performance. However, the tests used at interoperability events are not uniquely defined, so that one has to question on which basis

implementations are evaluated.

Web services, XML and SOAP

A Web service is a URL-addressable resource returning information in response to client requests. Web services are integrated into other applications or Web sites, even though they exist on other servers. So for example, a Web site providing quotes for car insurance could make requests behind the scenes to a Web service to get the estimated value of a particular car model and to another Web service to get the current interest rate.

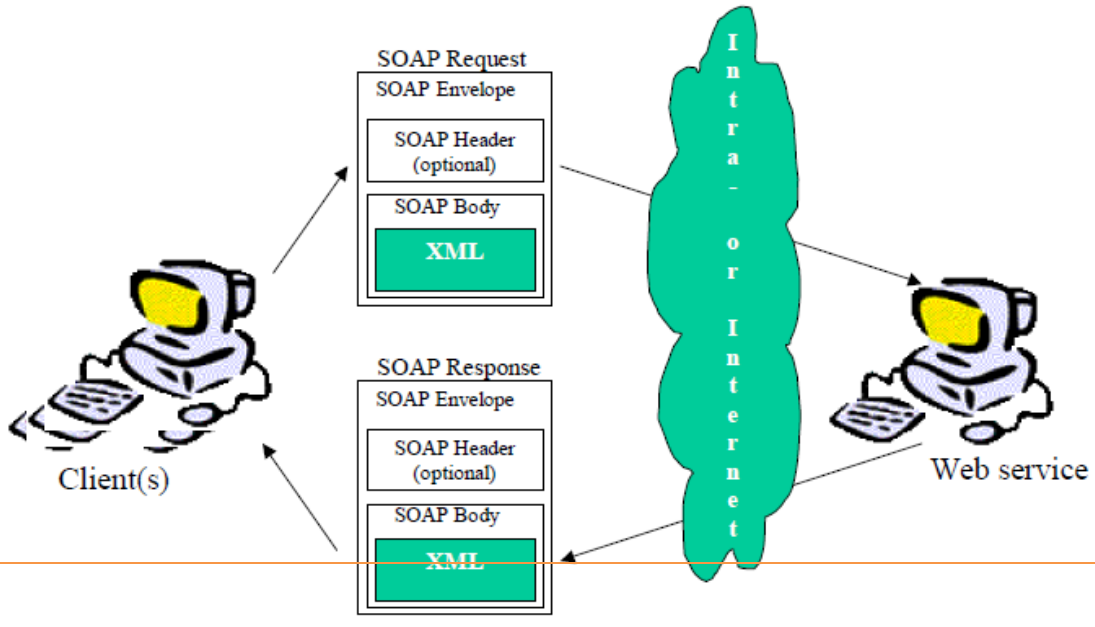


Figure 1. Principal Structure of a Web service

A Web service (see Figure 1) can be seen as a Web site that provides a programmatic interface using the communication protocol SOAP, the Simple Object Access Protocol: operations are called using HTTP and XML (SOAP) and results are returned using HTTP and XML (SOAP). The operations are described in XML with the Web Service Description Language (WSDL). Web services can be located via the Universal Description, Discovery and Integration (UDDI) based registry of services

XML DTDs and Schemas

XML stands for Extensible Markup Language and as its name indicates, the prime purpose of XML as for the marking up of documents. Marking up a document consist in wrapping specific portions of text in tags that convey a meaning and thus making it easier to locate them and also manipulating a document based on these tags or on their attributes. Attributes are special annotations associated to a tag that can be used to refine a search.

An XML document has with its tags and attributes a self-documenting property that has been rapidly considered for a number of other applications than document markup. This is the case of configuration files for software but also telecommunication applications for transferring control or application data like for example to Web pages.

XML follows a precise syntax and allows for checking well-formedness and conformance to a grammar using a Document Type Description (DTD).

SOAP

SOAP is a simple mechanism for exchanging structured and typed information between peers in a decentralized distributed environment using XML . SOAP as a new technology to support server-to-server communication competes with other distributed computing technologies including DCOM, Corba, RMI, and EDI. Its advantages are a light-weight implementation, simplicity, openstandards origins and platform independence. The protocol consists only of a single HTTP request and a corresponding response between a sender and a receiver but that can optionally follow a path of relays called nodes that each can play a role that is specified in the SOAP envelope. A Soap request is an HTTP POST request. The data part consists of:

- the SOAP envelope
- the SOAP binding framework
- the SOAP encoding rules

- the SOAP RPC representation called the body

Testing of Web services

Testing of Web services (as for any other technology or system) is useful to prevent late detection of errors (possibly by dissatisfied users), what typically requires complex and costly repairs. Testing enables the detection of errors and the evaluation and approval of system qualities beforehand. An automated test approach helps in particular to efficiently repeat tests whenever needed for new system releases in order to assure the fulfilment of established system features in the new release. First approaches towards automated testing with proprietary test solutions exist , however, with such tools one is bound to the specific tool and its features and capabilities.

Specification-based automated testing, where abstract test specifications independent of the concrete system to be tested and independent of the test platform are used, are superior to proprietary techniques: they improve the transparency of the test process, increase the objectiveness of the tests, and make test results comparable. This is mainly due to the fact that abstract test specifications are defined in an unambiguous, standardized notation, which Is easier to understand, document, communicate and to discuss. However, we go beyond "classical" approaches towards specification-based automated testing, which till now mainly concentrate on the automated test implementation and execution: we consider test generation aspects as well as the efficient reuse of test procedures in a hierarchy of tests.

Testing of Web services has to target three aspects: the discovery of Web services (i.e. UDDI being not considered here), the data format exchanged (i.e. WSDL), and request/response mechanisms (i.e. SOAP). The data format and request/response mechanisms can be tested within one test approach: by invoking requests and observing responses with test data representing valid and invalid data formats. Since a Web service is a remote application, which will be accessed by multiple users, not only functionality in terms of sequences of request/response and performance in terms of response time, but also scalability in terms of functionality and performance under load conditions matters. Therefore we have developed a hierarchy of test settings starting with separate functional tests for the individual services of a Web service, to a service interaction test checking the simultaneous request of different services, to a separate load tests for the individual services up to a combined load test for a mixture of requests for different services (see Figure 5). All the tests return not only a test verdict but also the response times for the individual requests.

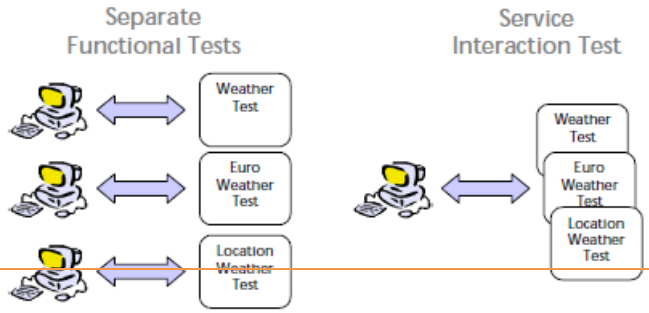


Fig: Test hierarchy for Web services

C.Mathematical Model

We model web service composition problem as a multi-criteria program. Especially, goal programming solves the problem in accordance with the customers' preferences and can find a compromised solution when the regular multi-criteria programming model does not have a solution. We

consider the customers' requirements on aggregating price, reliability of service, and total service time as goals. Both preemptive and non-preemptive goal programming models are built and tested.

Definition of Parameters and Variables of the Model

We define the entire list of variables and parameters in this section. Quality of Services (QoS) metric is used to evaluate services. In the context, cost, execution time and reliability are used to demonstrate the model

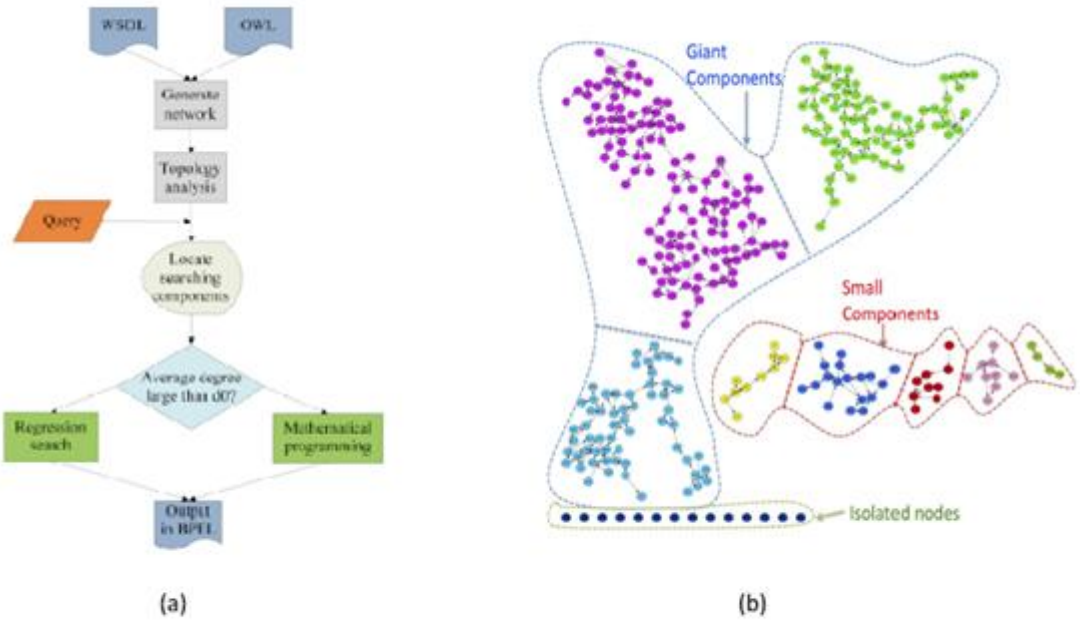


Figure Composition process (a) Composition framework, (b) An example of service networks

Hence, the QoS vector for a service can be expressed as:"quality of service(s) = f(cost(s), service execution time(s), reliability(s))." Cost is the expense for purchasing a service or services; service execution time is the process time of a service or services; reliability is a measure of successfully running a service or a set of services. The entire list of varibles used in the paper is shown in Table 1. In general, the number of attributes in the input set *I* and the number of attributes in the output set *O* are different. However, it is reasonable to let *n* = *maxfh; kg* be the total number of attributes since most of the attributes are the inputs of some services and the outputs of other services at the same time. Generally speaking, all the attributes can be inputs in some services and outputs in other services. Thus, it can be proved that *I* = *O* approximately, in large scale service networks. In order to define the

reliability score for web services, we give the following definition. The reliability of a service is a function of failure rate of the service.

If the failure rate of service *Z* is *f*, the reliability score of the service is defined as: $q(f) = log(f)$, where

$0 < f \_ 1$ and $0 \_ q(f) < +1$.

Here, we introduce reliability measure $q(f)$ in terms of the failure rate *f*. This technique is useful to convert the nonlinear objective function (7) into linear objective function (4), which simplifies the problem.

LP (linear programming) solvers can be used to solve the model. Next, we need to specify a depth level of composition before using this mathematical programming model. The decision about *L*, the depth level, is important as larger or smaller *L* influences the computational time and whether the optimal solution can be obtained or not.

Among all the variables we defined, the decision variables are Z_{lj}, the status of the j^{th} web service in the l^{th} level of composition , j=1,2,.....,m; l=1,2,.....,L.

$$Z_{ij} = \begin{cases} 1 & \text{web service } Z_j \text{ is selected in the } l^{th} \text{ level} \\ 0 & \text{otherwise} \end{cases}$$

j=1,2,.....,m; l=1,2,.....,L.

Objective Function

The objective function is defined as follows:

Cost:

The cost of the service composition equals to the sum of the prices of the services in the composition.

$$\min \sum_{l=1}^{L} \sum_{j=1}^{m} Z_{lj} \cdot p_j \qquad \text{..... (1)}$$

Definitions of variables and parameters:

Variable	Definition
Z	a set of web services
I	a set of input attributes of the web services
O	a set of output attributes of the web services
M	the number of services in Z
N	the number of attributes for the services in set Z
L	the maximal number of composition levels
Z_{lj}	web services that is currently available in the database; $Z_{lj} \in Z$; j= 1,2,….,m; l= 1,2,…..,L.
I_{lj}	the i^{th} input attribute of service Z_j ; i=1,2,….,n; j=1,2,…..,m.
O_{ij}	the i^{th} output attribute of service Z_j ; i=1,2,…….,n; j=1,2,…….,m.
p_j	the fixed price for acquiring the service from Z_j ; j=1,2,……,m.
t_j	the execution time of service from Z_j . j=1,2,……,m.
f_j	the failure rate of service Z_j . j=1,2,……,m.
q_j	the reliability of service Z_j . j=1,2,……..,m.

C_o	the maximum total cost that the customer is willing to pay for the services
T_o	the maximal total execution time that the customer allows to accomplish the entire process of the services
Q_o	the minimal reliability that the customer allows for a service in the composition
Q_1	the minimal overall reliability that the customer allows for the entire service complex , where $Q_1 > Q_o$

Service execution time:

The total process time for executing the entire series of services. We assume that the services at one level are executed in parallel.

The maximum execution time of the services in the 1st level is max { $t_j . Z_{1j}$ };

The maximum execution time of the services in the 2nd level is max { $t_j . Z_{2j}$ };

The maximum execution time of the services in the lth level is max { $t_j . Z_{lj}$ };

So, the total service execution time of this composition is:

$$\sum_{l=1}^{L} max_j\{ tj.Zlj \}$$

Let $_1$ be the maximum service execution time of the lth level. The above total service execution time expression can be reformulated in terms of following linear program:

$$\min \sum_{l=1}^{L} \eta_l \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

subject to $\eta_l - t_j * Z_{lj} >= 0.$ ..... (3)

j=1,2,.....,m;

l=1,2,.....,L.

Reliability:

Reliability of the service composition is described by the summation of the reliability scores of all the services included in the composition.

$$\max \sum_{l=1}^{L} \sum_{j=1}^{m} Z_{lj} \cdot q_j$$ ..... (4)

Constraints:

1.Input Constraints:

An input attribute of the query service should be included in the input attribute of the selected service in the in the composition. Thus,

$$\sum_{l=1}^{L} \sum_{j=1}^{m} I_{ij} \cdot Z_{lj} \geq I_{i0}$$ ..... (5)

i=1,2,.....,n.

2. Output Constraints:

An output attribute of the query service should be included in the output attribute of the selected service in the in the composition. Hence,

$$\sum_{l=1}^{L} \sum_{j=1}^{m} O_{ij} \cdot Z_{lj} \geq O_{i0} - I_{i0}$$ .....,. (6)

i=1, 2... n.

3.The reletionship of the output and inputs between the levels has to satisfy the following requirements. All the inputs of the selected services in the first level must be a subset of the initial set given in the query.

$$\sum_{j=1}^{m} I_{ij} \cdot Z_{1j} \leq I_{i0}$$ ..... (7)

i=1,2,.....,n.

Also , all the input sets of selected service at the k^{th} level must be a subset of the union of the initial input set given in the query and the output sets of services in previous level.

The formulation is:

$$\sum_{j=1}^{m} I_{ij} \cdot Z_{k+1,j} - \sum_{l=1}^{k} \sum_{j=1}^{m} O_{ij} \cdot Z_{lj} \leq I_{i0} \quad \text{..... (8)}$$

k=1,2,.....,L-1;

i=1,2,.....,n.

The reletion among the inputs of services in k^{th} level and the outputs from the previous levels and the attribute gives in the query needs to satisfy equation.

4. Goal constraint on the total cost:

The customer hopes that the total cost should not exceed C_0.

$$\sum_{l=1}^{L} \sum_{j=1}^{m} Z_{lj} \cdot p_j \leq C_0 \quad \text{..... (9)}$$

5. Constraint on the service execution time:

The customer hopes that the total service execution time should not exceed T_0 . Since some services can be executed in parallel , we take the longest execution time as the execution time of the set of services executed in parallel. The execution time of the composition, e.g. total service execution time is the sum of the service execution times of L levels. Thus,

$$\eta_1 >= Z_{lj}.t_j \quad \text{..... (10)}$$

D.Framework for Webservice Testing tool:

The application provided in this paper is the WSTT: Web Service Testing Tool. WSTT supports deep testing of Service – Oriented Architectures (SOA) consisting of any number of web services, which are tied to back-end databases.

- No-code SOAP/XML testing and WSDL exploration and test maintenance. All the information need to know is the URL to capture and invoke any type of test against a web service.
- WSTT is no-code automated testing, meaning no longer have to script tests.
- WSTT runs on any client and supports Java and any other SOAP compliant web service.

WSTT provide the way to perform several tests in very short period of time. It supports Functional Testing to ensure that the functionality of the web service is as expected. It allows to run all test cases one time for each web service. Figure 2 depicts the architecture of the proposed system.

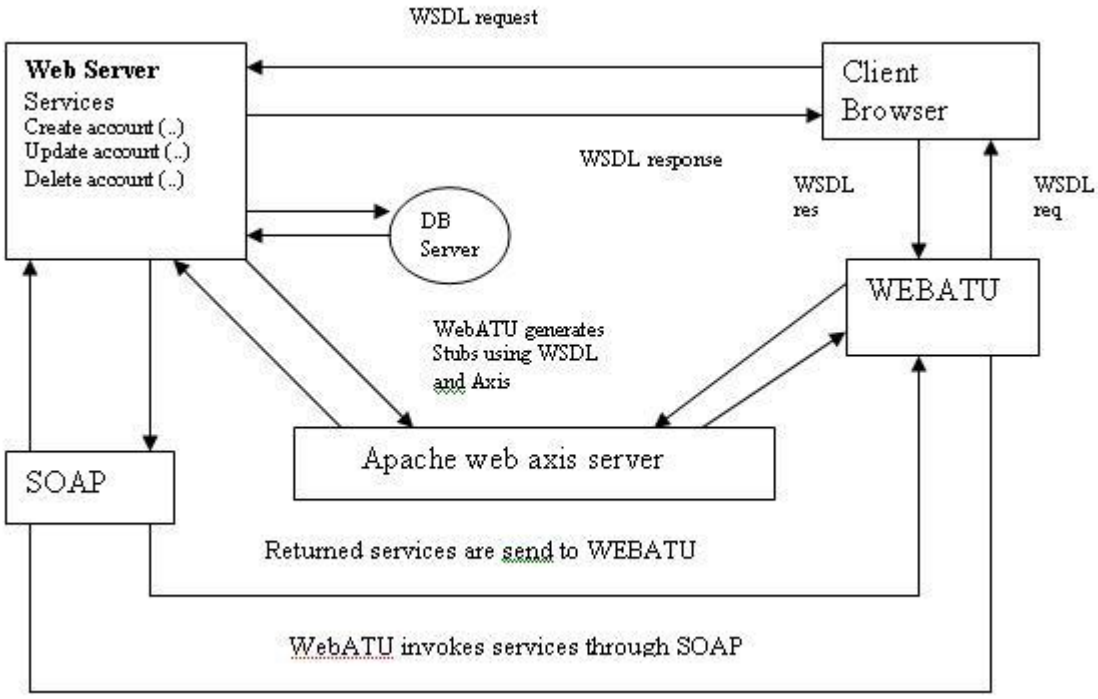


Fig: System Architecture

WSTT gives request to web server for WSDL through client browser, server gives WSDL file as response to client browser, and the tool fetch the wsdl from browser. With the help of Apache Axis client, WSTT generates all the required stubs to invoke and execute the available API's in the web services. WSTT communicates with server through SOAP messages which makes WSTT to be x language and platform independent.

The testing tool for web service interoperation is used for the interaction between users and services, it realizes the testing corresponding the standard of Web service interoperability by invasive black box testing method, can monitor the SOAP messages of web service in real-time, generates the monitoring statistic report including HTTP head, message content, etc. It also supports the analysis corresponding to the standard interoperability of UDDI registry, WSDL document, SOAP messages in web service, and then generates the consistency report about the interoperable problems in web service.

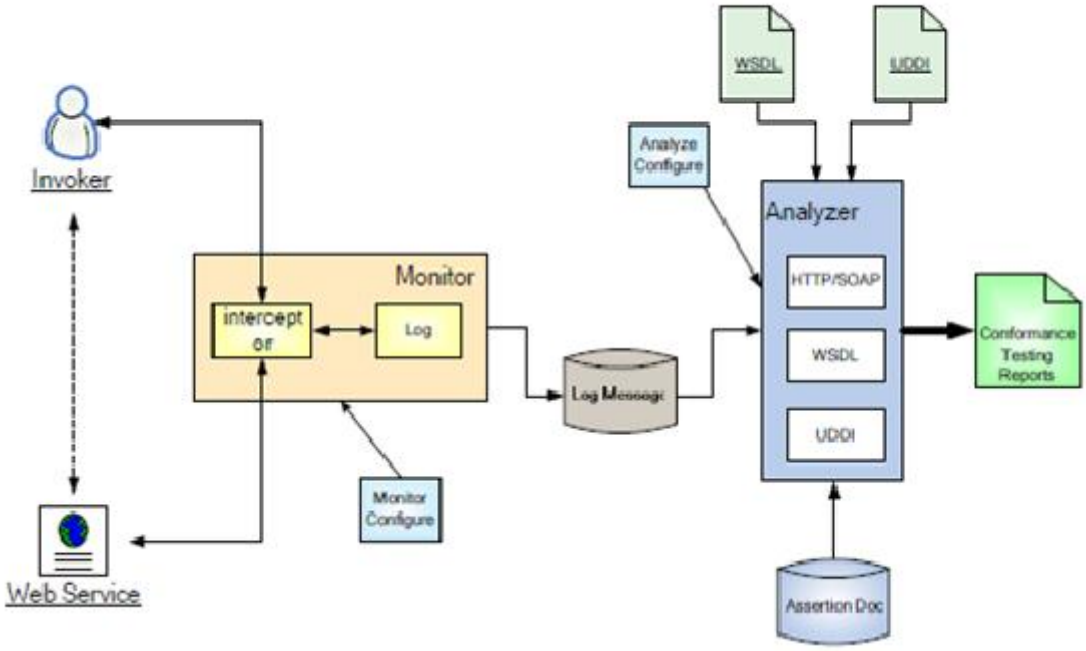


Fig.: Architecture of the testing framework

As shown in above figure, testing tool includes two modules: Monitor and Analyzer. Monitor intercepts the messages between service requester and web service in recessive method, transforms the interceptive message into standard format and then outputs it to message logs. Analyzer judges whether the components of web service meets the requirement of the interoperable standards using a suite testing assertions. As an independent

process, Monitor can start multiple socket threads at the same time, listen and relocate the server port, transmit and store the SOAP messages and eventually generates the log files. Monitor uses socket connection to get the service requesting response message, and monitors the SOAP communication between listener and client. The Host value is modified to listener port when the requesting SOAP messages are intercepted; the response SOAP messages have the same situation. Then monitor analyzes the connect type, message content, timestamp, coding method from the intercepted SOAP message and sends the results to log service. Finally, it send the SOAP message to original request point and response point, and ensures the normal proceed of business processes, which will realize the hidden information intercepting in a non-intrusive way.

Analyzer realized the interoperability consistency analysis based on testing assertions document, uses running testing assertion document as testing case. It can test and analyze UDDI registry, WSDL document, SOAP messages in web service, and then generates the consistency testing report. After the consistency analysis, analyzer generates the consistency report in fixed template and makes the initialization at first, then it calls the validateUDDI, validateWSDL, validateMessages functions to analysis various input sources, and writes the results to consistency report document. Finally, analyzer will make the conclusion that whether the last generation goes through the testing. The report contains the WSDL package about the important data of web service described in WSDL file (web service interfaces, data type information, binding information and address information), XML package used for XML document processing (reading and writing XML elements, generating XML document object model, processing XSL files), and provides the assist toolkit modules which have effective function support for other packages in the testing tool.

Web service solve the problems of traditional middle component technology, it makes the framework which is in different platforms or uses different object techniques removing the differences in realization and platforms, uniforms the frameworks in the same technical level, and the guarantee of its interoperability is the key point to call web service successfully. This document uses the basic outline of WS-I to analyze the web service interoperable problems. It designs and implements the testing tool corresponding the standard of web service interoperability which is based on the basic outline of WS-I, put forward the testing method corresponding to the standard of web service interoperability, design and implement the testing tool to test and analyze automatically. The next step of work is focused on making the interoperability problems haven't covered by the basic outline of WS-I as the extend points of the software testing, so as to expand the corresponding testing tool.

Framework:

This framework called WSTT for robustness testing of web services. Given a WSDL from a service provider, WSTT first generates code to facilitate both test generation and test execution. WSTT then generates a test suite, which includes tests with random or extreme method-argument values.WSTT runs the generated test suite on the generated client code, which eventually invokes the web service. WSTT then collects the results returned from the web service. In particular, our WSTT framework consists of code generation, test generation, test execution, and response analysis. The code generation component generates necessary code required to implement a service consumer. In addition, the component generates a wrapper class that can execute each service independently. The wrapper class contains the service calls. Unit tests are generated on this wrapper class to initiate SOAP requests to the service provider. The test generation component supplies the generated wrapper class to Java test generation tools in order to generate Unit tests. The test execution component executes the generated unit tests to cause the web service to be invoked and its responses to be collected. The response analysis component classifies and analyzes the collected responses from the web service. The component selects tests that cause the web service to return responses that are potentially robustness-problem-exposing.

Code Generation:

The code generation component generates Java clientside code from a service provider's WSDL. The client- side code is used to allow unit-test generation tools to automatically generate unit tests that exercise the various services offered by the service provider. WSDL is an XML-based language that describes the public interface of a service. It defines the protocol bindings, message formats, and supported operations that are required to interact with the web services listed in its directory. In particular, the component generates the following classes or interfaces for each WSDL file:

(1) A Java class is generated to encapsulate each supported message format for both the input and output parameters to and from the service. Message formats are typically described in terms of XML element information items and attribute information items.

(2) A Java interface is generated to represent each port type, which defines the connection point to a web service. A stub class is generated for each binding, which defines the message format and protocol details for a web service. The binding element has the name and type attributes. The name defines the name of the service, while the port of the binding is defined by the type attribute.

(3) A Service interface and corresponding implementation is generated for each service. The Service interface describes the exposed web services. This is a Java interface that facilitates web service invocation.

(4) A wrapper class is generated to allow to invoke the provided service. This wrapper class contains the web service SOAP call to the operations supported by the web service.

Test Generation:

The test generation component feeds the generated wrapper class to a Java unit-test generation tool to generate a test suite that exercises the services defined in the WSDL. The component operates relatively independently of test generation tools and thus any unit test generation tool for Java (such as JCrasher , Agitar Agitator , and Parasoft Jtest) may be used. Because the wrapper class drives the test generation tool, it is important that the wrapper class is constructed such that it leverages the strengths of the chosen generation tool. For example, if information regarding pre-conditions, legal service invocation sequences, and/or functional specifications is known, then it is possible to encode this information in the wrapper class to further guide the test generation tool in producing better tests. The component tries various possible combinations of several different special characters to form test inputs. For each non-primitive parameter, the component generates a null reference as a special argument value, which is eventually encoded by the SOAP engine via the xsi: nil attribute. In addition, the component constructs code fragments that will create instances of these complex data structures by passing different arguments to their Java constructors.

Test Execution:

Given the generated wrapper class, unit-test suite, and client-side implementation, the test execution component simply runs the generated tests. These tests invoke methods in the wrapper class. The methods in the wrapper class have been designed to leverage the generated client-side implementation to invoke the remote web service under test. Since the web service is remote, and faults and exceptions are expected to occur, we set a timeout parameter in the wrapper class of one minute for the execution of each test in the generated test suite. This timeout mechanism ensures that the test driver does not hang indefinitely during execution.

Response Analysis:

Manual inspection may be used to determine whether an exception should be considered to be caused by a bug in the web service implementation or the supplied inputs' violation of the service provider's preconditions. Even for the latter case, the web service implementation should respond with an informative error message rather than simply crashing information. For example, the GoogleSearchService has a string key argument that is used to uniquely identify the service requester. This identifier is received after registering with Google as a service consumer. If we allow this identifier to be automatically generated, then it is expected that the service should fail because the argument would violate the service provider's precondition, namely that the key identifier is valid. On the other hand, if the Google Search- Service fails due to a generated search string, it is likely an issue since this interface should be prepared to handle such cases. The response analysis component selects tests whose responses may indicate robustness problems and presents the selected tests for manual inspection. To collect web service responses, the component acts as a man-in-the-middle between the service consumer and the service provider. The service consumer directs the service request to the component, which records the request and forwards the request to the service provider. The component also records the service response or error condition returned by the service provider. Based on our experience of applying WSTT on various web services, we classify four main types of encountered exceptions that may indicate robustness problems:

1.404 File Not Found- The 404 or Not Found error message is an HTTP standard response code indicating that the client was able to communicate with the server, but the server either could not find what was requested, or it was configured not to fulfill the request and not to reveal the reason why.

2.405 Method Not Allowed- The HTTP protocol defines methods to indicate the action to be performed on the web server for the particular URL resource identified by the client. 405 errors can be traced to configuration of the web server and security governing access to the content of the site.

3.500 Internal Server Exception- In certain cases, the server fails to handle a specific type of generated random input and produces an Internal Server Exception with an error code of 500. This exception is the most common one and offers little insight into what the problem may be. *Hang.* The web service hangs indefinitely or the server takes more than one minute to return a response.

E.CONCLUSION

Due to its inherent complex nature, testing Web Services poses various challenges to test engineers. WSTT is developed after thorough study of Web Service Architecture and its related technologies. It is an excellent web service Functional Testing Tool. WSTT is very fast and execute with great speed

because it uses Sax Parsing. WSTT runs on any client and supports Java and any other SOAP complaint web service.

F. FUTURE WORK

The current features provide to amend the basic needs of tester and a programmer to test and verify functionalities of each API in the web service, irrespective of the platform and languages used in the development. However, elaborate results and reports on each API could be generated with enhancement of the current tool by embedding new features in the future. For example a new feature can be added as an enhancement that would generate report on each API after the completion of test execution, by which all the users can view the report on his/ her mail box. Also another feature can be added, which enables a user to run the test suit online, which will be more efficient and effective in the product testing life cycle. To eliminate the complexity associated with the changes that takes place in the WSDL, the location of the WSDL file can be referred to instead of saving the WSDL file in the local disk.

G. References

1. Elfriede Dustin. Effective software Testing-50 specific ways to improve your testing
2. William E.Perry. Effective Methods For Software Testing
3. Bosworth Adam;Developing Web Services;IEEE,2001
4. Curbera Francisco, Duftler Mathew, Khalaf Rania, Nagy William, Mukhi Nirmal,Weerawarana Sanjiva; Unraveling the Web Services:Web-An Introduction to SOAP,WSDL and UDDI;
5. IEEE Internet Computing,March/April 2002
6. Web Services Architecture, February 2004,www.w3.org/TR/ws-arch/
7. Web Services Description Language, August 2005,http://www.w3.org/TR/wsdl20/
8. SOAP Version 1.2 Part 0: Primer, June 2003,
9. http://www.w3.org/TR/2003/REC-soap12-part0-20030624/
10. Jeffery Buzen, Leo Parker, “Measurement Principles for Web Services”,December 2003
11. Mike Vermeulen, “Testing Web Services”, August 2002
12. Framework for Testing Web Applications using Selenium Testing tool with respect to Integration esting by 1Prasanth Yalla, Dr. L S S Reddy, 3M.Srinivas, 4T.Subha Mastan Rao,September 2011
13.Scenario Analysis of Web Service Composition based on Multi-Criteria Mathematical Programming By LiYing Cui and Soundar Kumara .

14.Automated Testing of XML/SOAP based Web Services Ina Schieferdecker, FOKUS, Berlin,Germany Bernard Stepien, University of Ottawa, Canada

IT 37

"An Approach to solve system Dynamics Problems using Fuzzy Logic"

[1]Mrs. Aparna B Kodgirwar
(Asst. Professor)
aparnakod@gmail.com
ASM'S Institute of Computer Studies,
Studies, Pimpari, Pune

[2]Mrs. Sheetal Deshmukh
(Asst. Professor)
deshmukh.mca@gmail.com
ASM'S Institute of Computer
Pimpri, Pune

Abstract:

System dynamics is a methodology and mathematical modeling technique for framing, understanding, and discussing complex issues and problems. It is an effective method for dealing with time varying and dynamic interactions among components in the complex system. It is generally used in domain of social , Economic and human activity systems which deals with imprecise and vague variables or events. In this kind of system casual loop is the main concept for model formulation. Some time casual loop can't be explained precisely and have vague and imprecise meaning. In some cases it is better to use system dynamics with the other models. Fuzzy logic is good interface for solving vague and imprecise problems in the system dynamics mode. This paper gives an approach for investigating fuzzy casual loop to study the behavior of fuzzy relations expressed by linguistic variables and present an alternative approach for analysis of the problems. This approach has used Expert System as well known tool in Artificial Intelligence for solving fuzzy Systems dynamics problems.

Keywords: Fuzzy set theory**,** Expert system, Fuzzy Graphs, Artificial Intelligence, Systems Dynamics.

1. Overview

Dynamic behavior is an essential feature of most complex systems. Complex means there are different types of variables and casual linkage. The influences of these variables on each other are also complicated. The direction of arrow shows the direction of cause effect relationship. The effect of one variable to another can be positive or negative or can be represented by linguistic variables. For example Low, High and very High and so on to use fuzzy set theory to simulate the effect.

i) Systems Dynamics

System Dynamics was developed by Forrester(1961-68) can be used to model and simulate complex systems. It is a methodology and mathematical modeling technique for framing, understanding, and discussing complex issues and problems. Also it have been used for nearly three decades to model economics. Social, human activity and other dynamic problems.

This technique is an effective methodology for dealing with time varying (Dynamic) interactions among components of analyzed system. A simulation language that is called DYNAMO was designed to simulate system dynamic model. The other application such as ITHINK and STELLA was designed for this purpose.

System dynamics technique identifies two sets of variables i) Level and ii)Rate

The Level variable consist of accumulation of resources within systems. The level equations describe the condition or the state of the system at any point in time. The rate variables represent a flow into or out of a level. Rate equation represents dependencies between flow rates and level.

The procedure of system dynamics modeling could be broken in to six major steps.

i) Problem identification and definition.
ii) System Conceptualization (Influence Diagramming).
iii) Model Formulation (Flow Diagramming, Equations).
iv) Simulation and Validation of the model.
v) Policy Analysis and improvement.
vi) Policy Implementation.

ii) Fuzzy set Theory:

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth- truth values between "completely true" and "completely false".

Fuzzy set theory has been studied extensively over the past 30 years. Most of the early interest in fuzzy set theory pertained to representing uncertainty in human cognitive processes. Fuzzy set theory was introduced by Zadhe(1965) is an idea to facilitate the use of uncertain concept in computation processes. Fuzzy number is a convex normalized fuzzy set that shows the fuzzy number imprecisely.

Computational linguistics is one of the most important applications of fuzzy sets. Fuzzy sets and linguistics variables can be used to quantify the meaning of natural language, which can be manipulate to solve the real problems. Linguistics variables are assigned values which are expressed such as word, phrases or sentences in a natural or artificial language. For example the value of part quality could be “high”, ”very high”, “low”, “not low” which are linguistics variables can be presented by fuzzy set. In this example “very’ and ‘not” are modifiers which use to modify our variables and describes our fuzzy concepts. A linguistics variables must have a valid syntax and semantic, which can be specified by fuzzy sets and rules. As already explained, influence or casual loop diagramming is the most important step in system dynamics modeling. The casual loop diagram explains how the system works. This is created through writing the names of the variables and connecting them by an arrow or link. The directions of the arrow show the direction of causation. This set is a signed directed graph and if a link expresses fuzzy relation. This graph is fuzzy signed directed graph (Fuzzy-SDG). Fuzzy graphs are graphs that its relations are fuzzy. In following figure fuzzy graph is represented. The relations of this graph are between numbers 0 and 1.

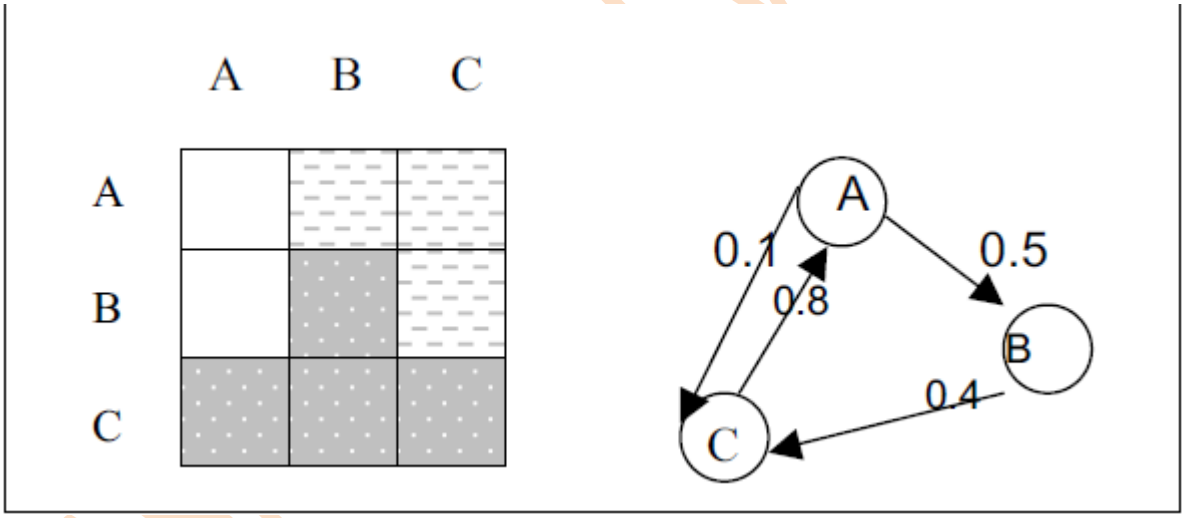


Figure 1: Fuzzy Graph

The mathematical notation for a graph is defined as below:

$G(x_i , x_j) = \{(x_i , x_j), \mu_G(x_i , x_j) \mid (x_i , x_j) \in E \times E\}$

In this example the notation is represented here:

Fuzzy_ G = {((A,B),0.5),((B,C),0.4),(C,A),0.8),((A,C),0.1) }

Fuzzy graph is a notation that is used in fuzzy system dynamics model.

iii) Expert Systems

In artificial intelligence, an expert system is a computer system that emulates the decision-making ability of a human expert. Expert systems are designed to solve complex problems by reasoning about knowledge, like an expert, and not by following the procedure of a developer as is the case in conventional programming. An expert is a person who is expertise in a certain area. The elements of typical expert system are: Knowledge based (Rule), Inference engine (Agenda), working memory (Facts), Explanation facility, and knowledge acquisition facility and user interface.

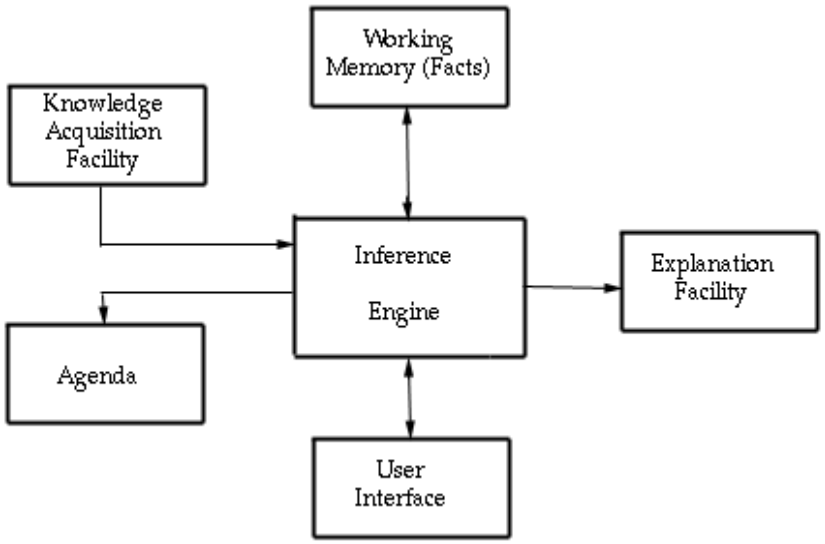


Figure2: Components of Expert System

In a rule based expert system the needed knowledge for solving problems is coded in the form of rules. Other types expert system used different representations for knowledge. Rules are conditional type statements that comprise of an antecedent and a consequent port ion. The form of rule is:

If antecedent **Then** consequent

For example: **If** quality is very low **Then** cost of repair became very high.

Fuzzy rule is a kind of rule that its element is fuzzy or linguistic variables. Much fuzzy knowledge exists in the real word, i.e. knowledge that is vague, imprecise, uncertain or ambiguous in nature. In crisp system, which based upon classical theory it is very difficult to answer some questions because either they do not have true answer completely, or in many times it is not possible to state the condition precisely. For example a manager may not able to state precisely the condition that under it "quality of part" affects on repair cost of part" but he/she can state that:

'**If** quality of part is low **Then** repair cost of part become very high."

Thus use of fuzzy system help us to solve this problem and cope with this kind of information.

2. System Dynamics with fuzzy logic

A system dynamics model can be consider to have fuzzy logic if it has at least one of the character below

i) Some of the level variables are fuzzy.

ii) The time agent may be vague.

iii) Some of the rate variables are fuzzy.

iv) Some of the auxiliary variables and the other variables are fuzzy.

v) Some of the relations can be replaced by conditional statements, which include fuzzy variables. This conditional statement generally follow an IF- Then format.

vi) Some of the relations can be presented by fuzzy algorithm, which include fuzzy variables. Fuzzy algorithm is a fuzzy is a fuzzy statement, which can be represented by IF-DO format.

If the cost of Inventory is "very low" Do the Total Cost "Very low"

vii) The degree of uncertainty of variables when the available information regarding them is imprecise or incomplete can be represent by fuzzy probabilities(likely, unlikely)

Some of the operator may be fuzzy. According to these characteristics we proposed an Approach to solve system Dynamics Problems using Fuzzy Logic that will explain in the next part.

3. A new Approach for Fuzzy System Dynamics

a) Steps of Approach

The proposed Approach has following steps:

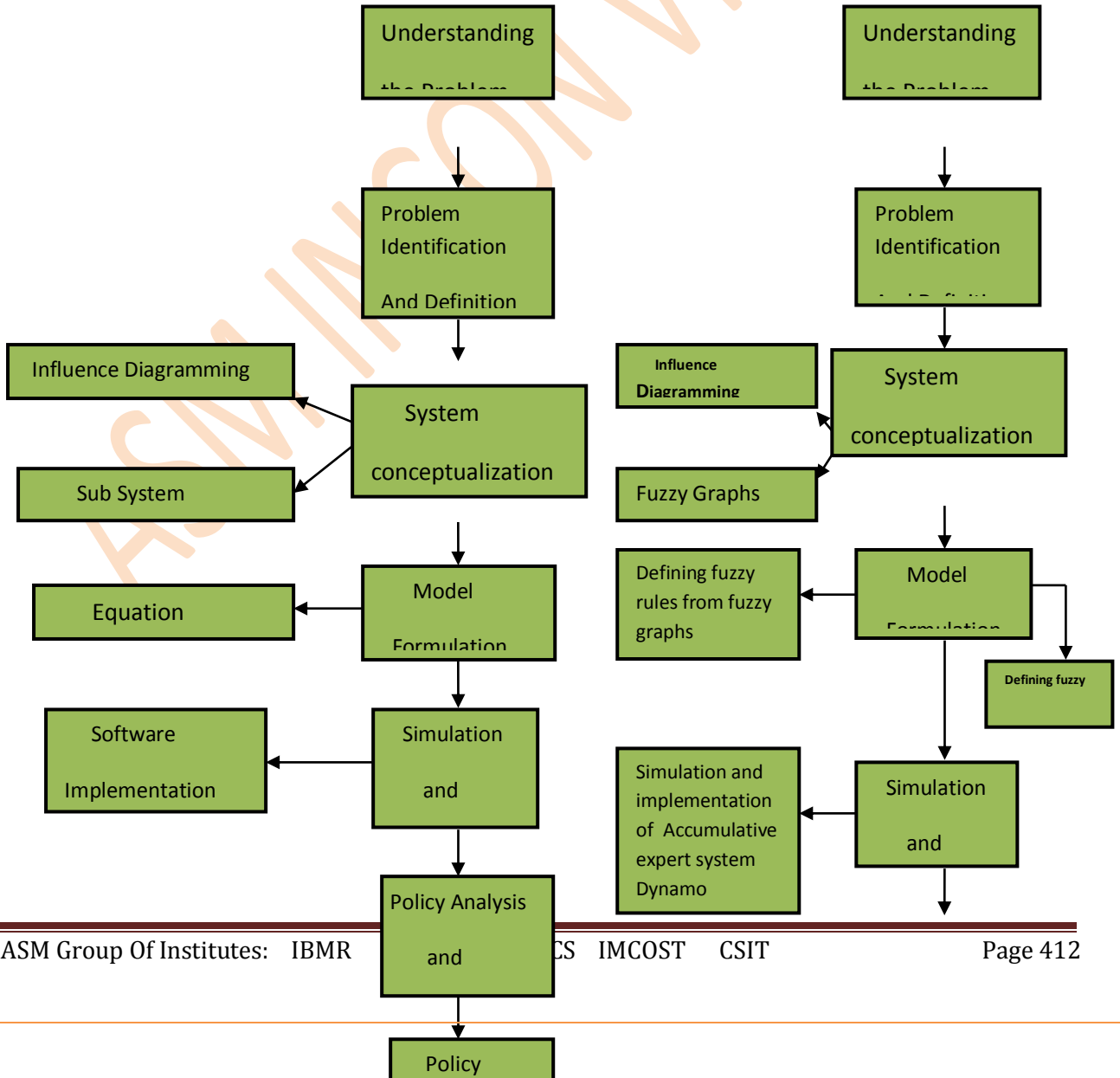
i) Description of problem
ii) Collecting related data
iii) Definition of imprecise data and linguistic variables
iv) Definition of the effects of the system component on each other and show theieffects in the form of fuzzy graph.
v) Translation of the fuzzy graphs into fuzzy rule with the IF-Then form.
vi) Definition of uncertainty of each rule.
vii) Creating fuzzy facts from the natural behavior of the system that express by the experts.
viii) Implementing a fuzzy expert system via the production rule and facts.
ix) Analysis of expert system outputs(The output can be either fuzzy variables or a defuzzified numbers.)

The proposed approach is a solution for fuzzy system dynamics. In this model we combine fuzzy graphs, fuzzy expert system many of usual steps of

system dynamic models and at the end of this steps we implemented a fuzzy expert system by means of a shell(Fuzzy Clips).

b) Model Validation Steps of this new methodology are similar to the usual steps of the system dynamics methods. But in this methodology we use Expert system as a known tool and our method is complex of the system dynamics property and expert system concepts.The below figures compares the new methodology and the system Dynamics Methodology.

We add new properties to an expert system in order to implement system dynamics Properties. For example when we add accumulation property in our model it is better to save the last important data and add the last value to the new value for saving information. After firing a rule the old value adds to the new value. This effect help us to model the accumulation property.



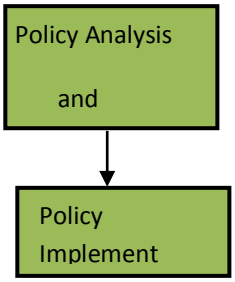


Fig1: Steps of system Dynamics **Fig2: Steps of New Approach**

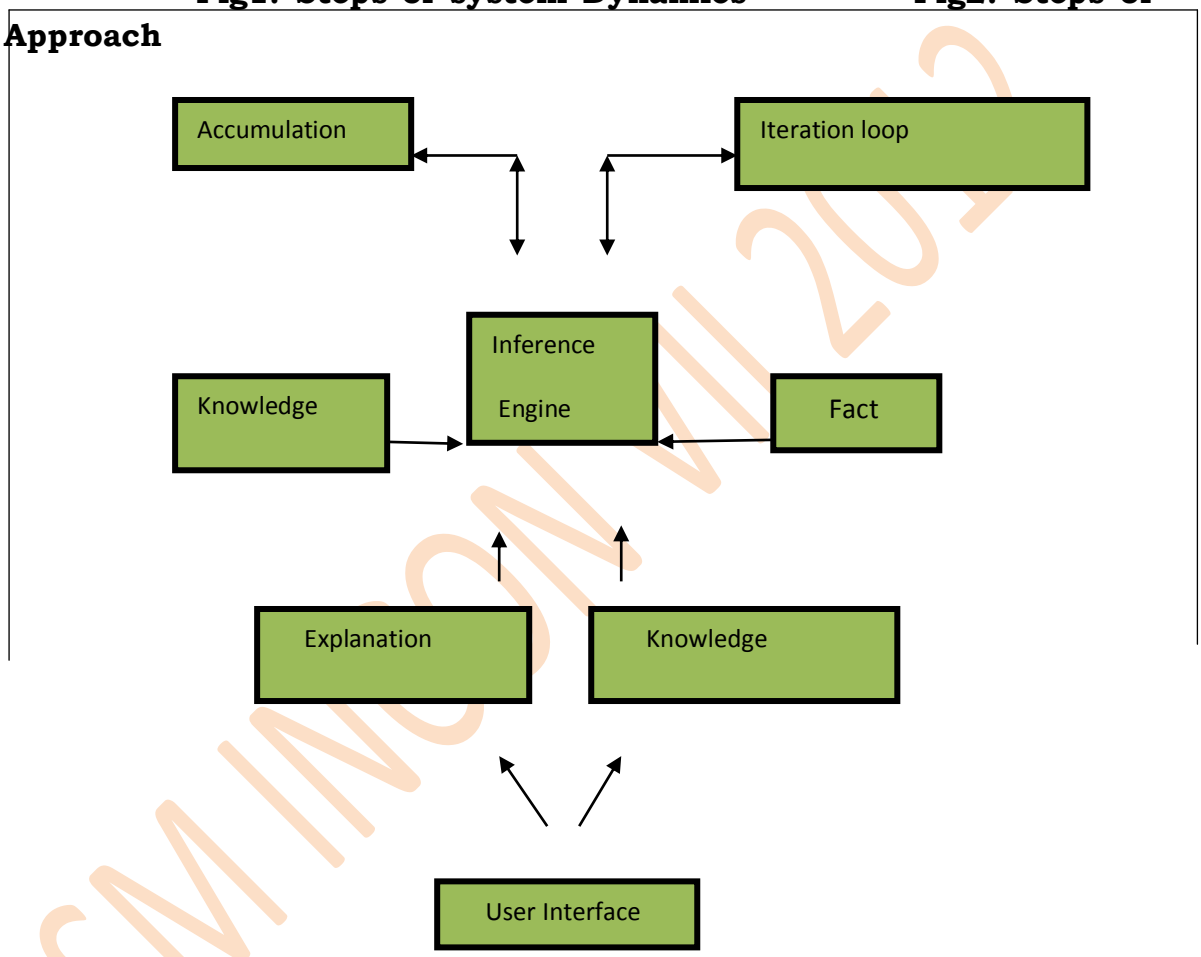


Fig 3: A new changed expert system components for new model

The above figure shows these new components that are added to expert system. There is two alternative which we can use them to implement this property.

a) Using modifier is the first alternative. “very” ,”little’ and so on are sample of this modifiers. We can use them to implement these modifiers to calculate a value as an accumulative value.

Value (I+1) =Value (I) + Very High (Value (I+1))

b) Deffuzifing is the second alternative This alternative adds old equivalent deffuzzy data to new equivalent deffuzzy data. This adding will be considered.

Deffuzified (value (I+1)) = Value (I) + Deffuzified (Value(I+1))

Following figure shows these two ways how two way accumulations works:

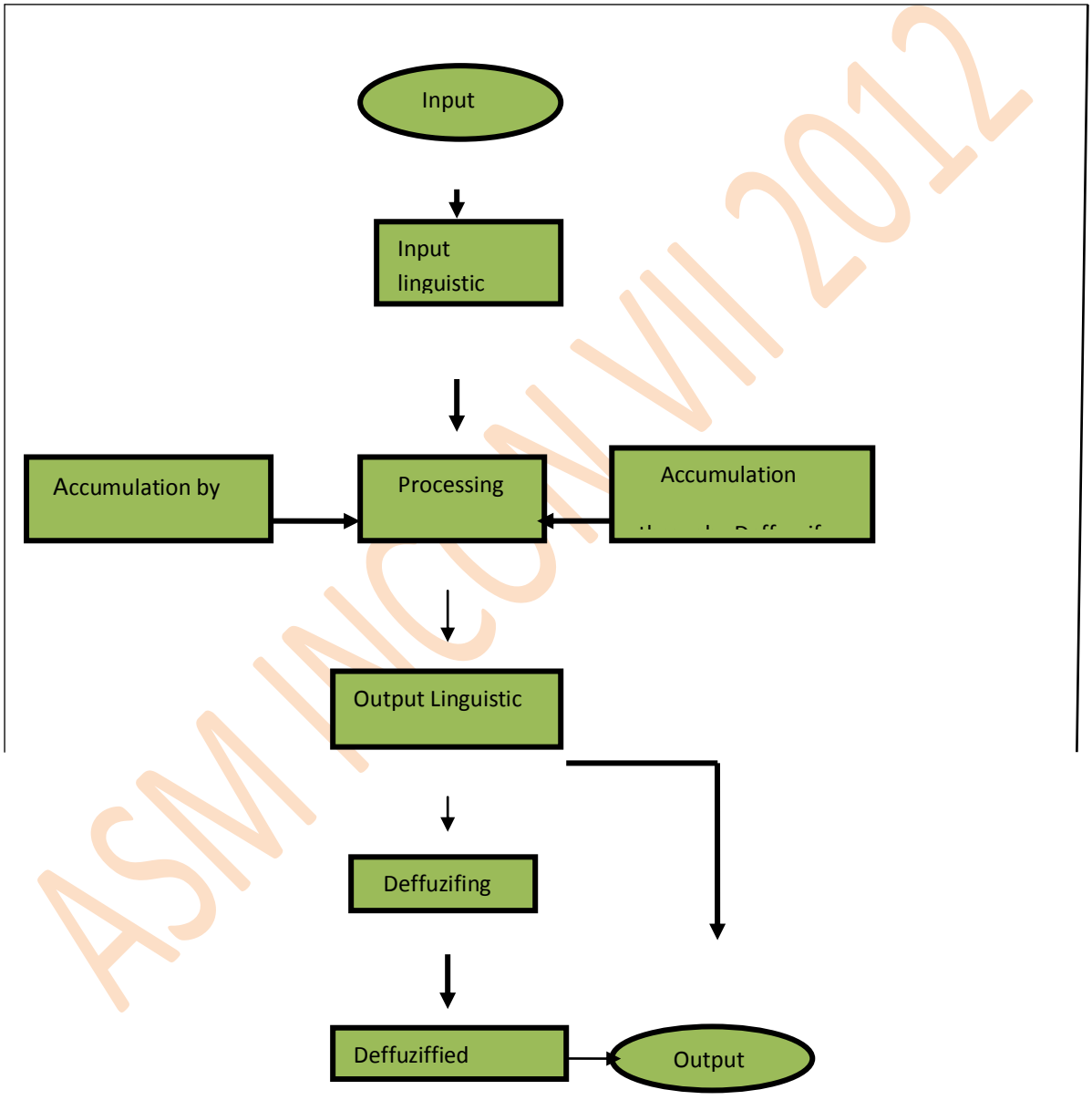


Fig 4: Two way of the accumulation

4. Illustration

Example of system Dynamics, which include fuzzy and linguistic variables, are described below. This proposed method has been illustrated through an example.

The following are the steps of model:

i) Description of problem
ii) Collecting related data.

This example is about a purchase system, which consists of following parameters:

Lead Time(LT): This is time, which is interval in that purchase must be done.

Supplier number(Supp_No): It is number of suppliers, which have the acceptable parameters for our purchase. For example their quality must be good and their cost must be low.

Quality(Q): It is quality of the parts of which we purchase from the suppliers.

Cost of parts(Cpart): It is cost of parts.

Cost of repair(CR): It is repair cost and related to quality of parts.

Cost of Inventory(C_inv): It is part of cost that exist in the inventory.

Cost of Line Sleep(C_sleep): It is the cost of sleeping the production line.

Total cost(C_total): It is total cost that influence of inventory cost.

ii) Definition of imprecise data and linguistic variable.

IF Q equal to low **THEN** CR with certainty factor 80 percent becomes low.

IF LT equal to low **THEN** Cinv CR with certainty factor 80 percent becomes low.

IF LT equal to low **THEN** Csleep CR with certainty factor 80 percent becomes high.

IF Supp_No equal to low **THEN** Cpart CR with certainty factor 80 percent becomes high.

IF CR equal to low **THEN** C_Total CR with certainty factor 80 percent becomes low.

IF Cpart equal to low **THEN** C_Total CR with certainty factor 80 percent becomes low.

IF Cinv equal to low **THEN** C_Total CR with certainty factor 80 percent becomes low.

IF Csleep equal to low **THEN** C_Total CR with certainty factor 80 percent becomes low.

iii) Definition of the effects of the system component on each other

Figure 2 shows the effects of system component on each other in the form of fuzzy graph. This graph exhibits the relation between components of this purchase system and also shows the fuzzy parameters and linguistic variables.

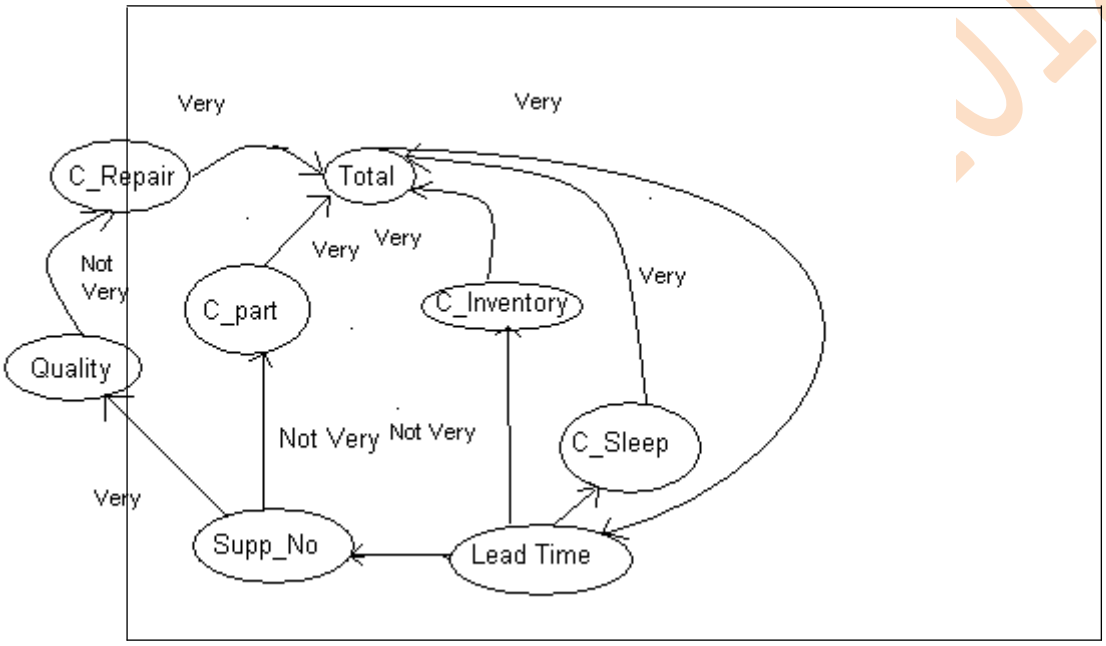


Figure5: Fuzzy graph for purchase system

Fuzzy_G = {(Q,CR),very),((Supp_No,Q),Very),((LT,Supp_No),very), ((

LT,Cinv),very),((LT,C_Sleep),not very),((Supp_No,Cpart),not very),

((CR,C_Total),Very),((Cpart,C_Total),ery),((Cinv,C_Total),very),

((Csleep,C_Total),very)}

iv) Translation of Fuzzy Graphs into fuzzy Rules with the IF_THEN form.

v) Definition of uncertainty of each Rule.

IF Q == low THEN CR = high

IF Supp_no == low THEN Q = low

IF LT == low THEN Supp_No = low

IF LT == low THEN Cinv = low

IF LT == low THEN Csleep = high

IF Supp_No == low THEN Cpart = high

IF CR == low THEN C_Total = low

IF Cpart == low THEN C_Total = low

IF Cinv == low THEN C_Total = low

IF Csleep == low THEN C_Total = low

vi) Creating fuzzy facts from the natural behavior of the system that express by experts.

vii) Implementing a fuzzy expert system via this production rules and facts.implementation an expert system from these rules, accomplished by fuzzy CLIPS. By using of this software, we can define linguistic variable and define fuzzy rules.

viii) Analysis of the expert system outputs

This output can be either fuzzy variables or a defuzzified numbers. For developing system dynamics accumulative attributes, System use accumulative variables to simulate this attribute, in each loop we add present data of this accumulative variable to old data. Implementation of this accumulative has done by two ways.

a) By using modifiers we can accumulate fuzzy variables. For example "high" linguistic variables after accumulation is changed to "very high', "very very high" and "extremely high". This method(accumulation of linguistic variables) help us

To change "Static Expert System" to Dynamic expert system" and lets to simulate this system for the desired steps. For example we can simulate this purchase expert system for 100 times.

b) In second way we deffuzified the fuzzy numbers and add their values with their old values and during the simulation we transfer their data to a temporary file. At the end of simulation we transfer this file to MATLAB software and plot the graph of each parameter. from the graph we can see the behavior of parameter.

5. Conclusion

In real life, because of the uncertain information as well as the vague human feeling and recognition, it is difficult to make an exact evaluation in social problems. Social system, economic system and and every system that

deal with human parameters have imprecise behavior and no one can state those parameters precisely. Therefore fuzzy logic and linguistic variable can help us in this types of systems.

In this paper we have discussed the use the Fuzzy expert system to represent uncertainty in system dynamics models. The proposed method can be used for investigating the system dynamics under vague and ambiguity conditions and is useful for the intelligent control over the dynamic systems. The proposed method can be used to implement in some of the applications like, Transportation problems, Assignment Problems, ATM systems etc just as shown in the above illustrated example of inventory system.

References

1. **"System Dynamics with fuzzy logic",** Levary R.R ,Int.J.System Sci,Vol.21.
2. **System Dynamics a practical approach for managerial problems**, Sushil
3. **Fuzzy Sets and Its applications:** Zimmerman,Kluwer-Nijhoff Publishing
4. **Expert Systems principles and programming,** Giarratano J,PWS Publishing Company
5. **"Comparison of fuzzy and crisp system via system dynamics simulation",** Polat S.,European journal of operational research,Vol 138.
6. **"Fuzzy System Dynamics: an approch to vague and qualitativevariables in Simulation",**Tessem B, and Pal I.Davisen, System Dynamics Review.Vol 10
7. **"A fuzzy set theoretic approch to qualitative analysis of casual loops in system Dynamics".** Pankaj, European journal of operational research,Vol 7
8. **"Fuzzy dynamics in software project simulation and support"**, Lehman, the working paper in natural science and Technology and Medicine.

IT 038

Comparison of Various Filters Applied on Different Types of Noises in Images under Image Processing Techniques

Name: Divya Khandelwal

Designation: Assistant Professor

Email-id: gargdivs22@gmail.com,

Name: Pallavi Gillurkar

Designation: Assistant Professor

Email-id: pallavi.2106@gmail.com

Contact No. 9579657701

Abstract

Image Processing is a technique in which the data from an image are digitized and various mathematical operations are applied to the data, generally with a digital computer, in order to create an enhanced image that is more useful or pleasing to a human observer, or to perform some of the interpretation and recognition tasks usually performed by humans. Image processing is used to solve identification problems, such as in forensic medicine or in creating weather maps from satellite pictures. A common inverse problem in image processing is the estimation of an image given a corrupted version. This problem is generally known as image restoration. The purpose of image restoration is to undo the defects which degrade an image. In this paper the problem in images degraded by camera movement as well as by object movement is examined. We have also discussed various applications of Image restoration and various filters which can be applied to poor quality images to either remove noise or improve the visual appearance. Moreover, the filters can be applied iteratively. Experiments with various filters show that the Winer Filter is the best among all other filters to improve the image quality.

Key Words: Image Processing, Image Restoration, Degradation, Filters

1. Introduction

As we are all aware of the fact that information carries the prime importance in today's rapidly growing technological advancement where anything that supports information is viewed very seriously. This means that all of us want very-very clear and accurate information at opportune moments to get the insure quality output, and the importance of colour information in digital image processing is greater than ever. Interest in digital image processing methods stems from two principal application areas: improvement of pictorial information for human interpretation; and processing of image data for storage, transmission, and representation for autonomous machine perception.

This paper proposes various Image processing steps along with various applications of Image restoration and various filters which can be applied to poor quality images to either remove noise or improve the visual appearance. Experiments with various filters show that the Wiener Filter is the best among all other filters to improve the image quality.

2. What is an Image?

An image could be an analog or digital. Analog images represent the visual world continuously. Photographs produced on film are thought of as analog, even though in the limit they are not. Digital images represent the world discretely.

An image may be defined as a two-dimensional function, f(x,y), where x and y are spatial (plane) coordinates, and the amplitude of f at any pair of coordinates (x,y) is called the intensity or gray level of the image at that point. When x, y, and the amplitude values of f are all finite, discrete quantities, we call the image a digital image. Digital image is composed of a finite number of elements or pixels; arrays of different light levels, represented in the computer by a number.

3 Image Representations

Images are usually represented in matrix form. In image representation, one is concerned about the characterization of quantity that each picture element represents. Important consideration in image representation is-

- Fidelity or intelligibility criteria for measuring quality of image

- Performance of a processing technique.

Specifications of such measures require models of perception of contrast, colour and spatial frequencies. Knowledge of fidelity helps in designing the imaging sensor as it provides information regarding the parameters to be

measured accurately.

A digital image is an array of real or complex numbers represented by a finite number of bits, as shown below in Fig. 1.

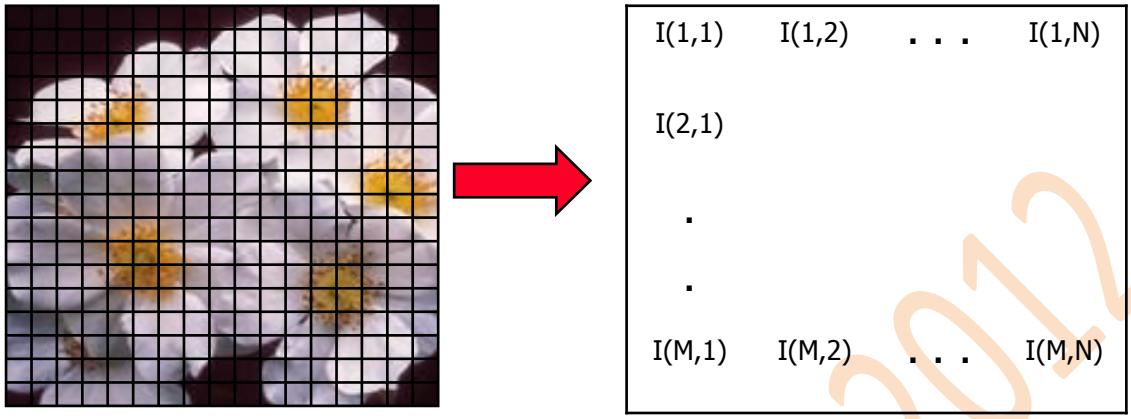


Fig. 1 An array of digital image

4 Image Enhancement

The objective of image enhancement is to accentuate certain features of the image for subsequent analysis or simple display so that an improved image quality can be obtained. Some of the common image enhancement techniques are:

- Contrast & edge enhancement
- Pseudo colouring
- Noise filtering
- Sharpening of image
- Magnification of images

Image enhancement is useful in feature extraction, image analysis and visual information display. It simply emphasizes certain image characteristics and the algorithms are interactive in nature. Enhancement processes increase the dynamic range of the chosen features so that they can be detected easily.

Thus, the image enhancement is carried out due to poor resolution in images, poor contrast due to under or over-exposure, corruption with noise. Fig 2 shows the enhancement of image of a car number plate. One can see the difference in the original image and the image after enhancement.

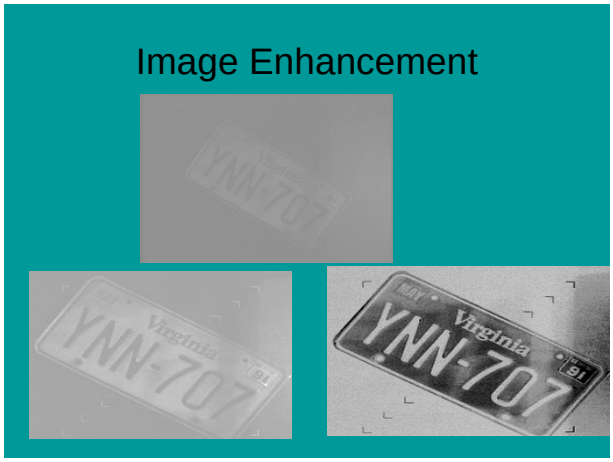


Fig 2: Enhancement of image of a car number plate

5 Image Restoration

Any image acquired by optical, electro-optical or electronic means is likely to be degraded by the sensing environment. The degradations may be in the form of sensor noise, blur due to camera misfocus, relative object-camera motion, random atmospheric turbulence, and so on. Image restoration is concerned with filtering the observed image to minimize the effect of degradations. The effectiveness of image restoration filters depends on the extent and the accuracy of the knowledge of the degradation processes as well as on the filter design criterion.

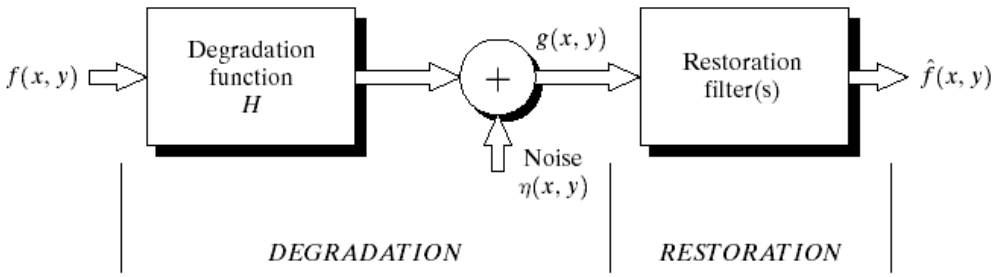


Fig 3: Image degradation/restoration process model

5.1 Image Noise

Quite often an image gets corrupted by noise, which may arise in the process of acquiring the image, or during its transmission or even during reproduction of the image. Removal of noise from an image is one of the most important tasks in image processing. Depending on the nature of noise, such as additive or multiplicative type of noise, there are several approaches towards removing noise from an image. The knowledge about the imaging system and the visual perception of the image helps in generating the noise models and estimating of the statistical characteristics of noise embedded in an image is important because it helps in separating the noise from the useful image signals. We describe two important classes of noise here:

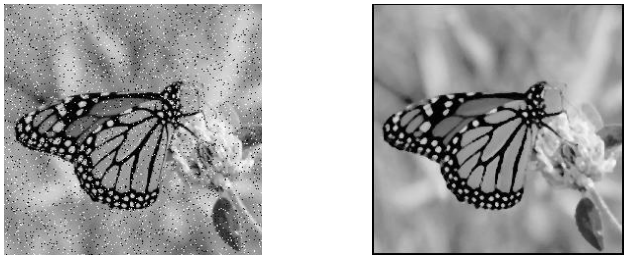
1) **Addictive noise:** Sometimes the noises generated from sensors are thermal white Gaussian, which is essentially addictive and signal idependent, i.e g(x,y) = f(x,y)+ n(x,y), where g(x,y) is the result of the original image function f(x,y) corrupted by the addictive Gaussian noise n(x,y).

2) **Salt-and-Pepper Noise (Impulse Noise)** Salt and pepper noise is sometimes called impulse noise or spike noise or random noise or independent noise. In salt and pepper noise (sparse light and dark disturbances), pixels in the image are very different in color or intensity unlike their surrounding pixels. Salt and pepper degradation can be caused by sharp and sudden disturbance in the image signal. Generally this type of noise will only affect a small number of image pixels. When viewed, the image contains dark and white dots, hence the term salt and pepper noise [5]. Typical sources include flecks of dust inside the camera and overheated or faulty (Charge-coupled device) CCD elements. An image containing salt-and-pepper noise will have dark pixels in bright regions and vice versa. This type of noise can be caused by dead pixels, analog-to digital converter errors and bit errors in transmission.

5.2 Some of the commonly used filters are:

1. Median filter
2. Laplacian filter
3. Inverse Filter
4. Wiener filter

Median filter: In image processing it is usually necessary to perform high degree of noise reduction in an image before performing higher-level processing steps, such as edge detection. The median filter is a non-linear digital filtering technique, often used to remove noise from images or other signals. It is particularly useful to reduce salt and pepper noise. Its edge-preserving nature makes it useful in cases where edge blurring is undesirable.



(a) (b)

Fig 4: (a) Original image with salt and pepper noise. (b) Median filtered image

Laplacian filter: The Laplacian filter is used for detection of edges in an image. It highlights areas in which intensity changes rapidly producing a picture of all the edges in an image. It enhances the details in all directions equally.

(a)

(b) (c)

Fig 5: (a) original image. (b) Laplacian filtered image. (c) Contrast enhanced version of Laplacian filtered image.

Inverse Filter: To deblur the image, we need a mathematical description of how it was blurred. (If that's not available, there are algorithms to estimate the blur.) We usually start with a shift-invariant model, meaning that every point in the original image spreads out the same way in forming the blurry

image. We model this with convolution: G(x,y)=F(x,y) x H(x,y) , where F is the image function, H is the complex distorting function, G is the "observed" image function. Now the main problem is to we need to recover F(x,y) from the convolution G(x,y)=F(x,y)*H(x,y).

The solution to this problem is to filter by dividing the blurred function by the reciprocal of H(x,y):G(x,y)/H(x,y)=F(x,y) x H(x,y)/H(x,y) = F(x,y).This is commonly referred to as the inverse filtering method where 1/H(x,y) is the inverse filter.

Difficulties with Inverse Filtering:

The first problem in this formulation is that 1/H(x,y) does not necessarily exist. If H(x,y)=0 or is close to zero, it may not be computationally possible to compute 1/H(x,y). If there are few values of H(x,y) which are close to zero then the ideal inverse filter can be approximated with a stabilized version of 1/H(u,v) given by : **Fapprox(x,y)=G(x,y) x Hinv(x,y)** where **Hinv(x,y)= 1/H(x,y) if |H(x,y)| > threshold value**

= 0 otherwise

This works well if few elements of H have a magnitude below the threshold but if two many elements are lost, the frequency content of Fapprox will be much lower than F(x,y) and the image will appear distorted.

Another problem with inverse filtering is that it doesn't perform well when used on noisy images. We revise the "observed image" model by including an error term:

Gn(x,y)=G(x,y) + N(x,y).

Wiener filter: The most important technique for removal of blur in images due to linear motion or unfocussed optics is the Wiener filter. The goal of the Wiener filter is to filter out noise that has corrupted a signal. It is based on a statistical approach. One can easily see the difference between Inverse Filtering and Wiener Filtering when applied on distorted image, in fig 6, given below.

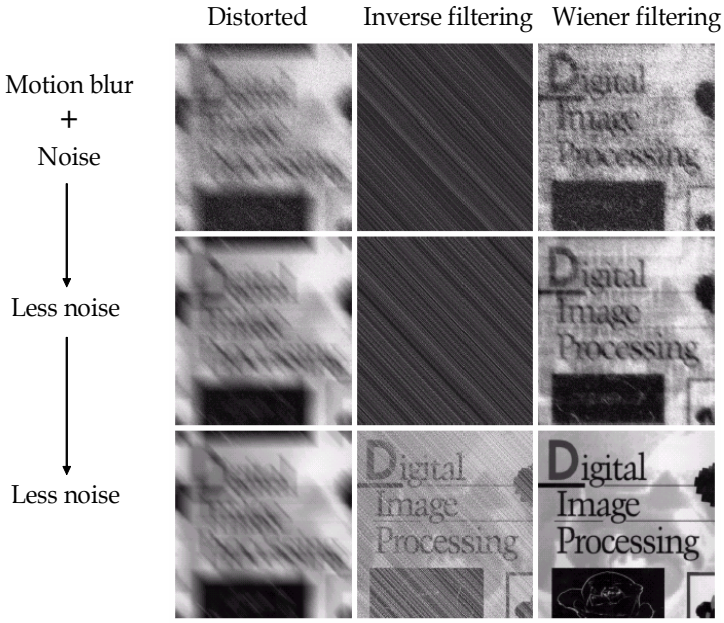


Fig 6: Inverse vs. Wiener Filtering

5.3 Image reconstruction

An important problem in image processing is to reconstruct a cross section of an object from several images of its transaxial projections. Image reconstruction from projections is a special class of image restorations problems where a two-(or higher) dimensional object is reconstructed from several one-dimensional projections. A projection is a shadow gram obtained by illuminating an object by penetrating radiations. Each projection is obtained by projecting a parallel X-ray beam through the object. Planar projections are thus obtained by viewing the object from many different angles. Reconstruction algorithms derive an image of a thin axial slice of the object, giving an inside view. Such techniques are important in medical imaging (CT scanners).astronomy, radar imaging, geological exploration, and nondestructive testing of assemblies. Figures 7, 8 and 9 show an excellent example of image reconstruction.

STANDING BUDDHA STATUE IN BAMIYAN N-W OF KABUL, AFGHANISTHAN CARVED OUT OF THE ROCK

Fig 7:

BUDDHA IMAGES DEMOLISED ON MARCH 2001

Fig 8:

IMAGES ON INTERNET

3 D MODEL USED FOR RECONSTRUCTION

Fig 9:

6. Limitations

The selection of the denoising technique is application dependent. So, it is necessary to learn and compare denoising techniques to select the technique that is important for the application in which we are interested. Also the following points needs to be considered while handling with images.

- Huge Data Storage
- Software needed to handle the images which are expensive
- Experienced and Trained Manpower required
- increased computational complexity

7. Conclusion

In this paper, we discussed different filtering techniques for removing noises in digital images. Furthermore, we presented and compared results for these filtering techniques. The results obtained using different filter technique ensures noise free and good quality of the image. Also by applying various

filters on different types of noises we find the performance of the Wiener Filter after de-noising for Gaussian noise and Salt and peeper noise is better than Inverse Filter But this technique increases the computational complexity. The main advantages of the Winer filter are the denoising capability as compared to other Filters which have been discussed in this paper.

8. References

1. Andrews, H.C. and Hunt, B.R., Digital Image Restoration, Pretence Hall, Englewood Cliffs, New Jersey, 1977.

2. B. Chanda and D. Dutta Majumder (May 2006), Digital Image Processing and Analysis, 2nd Edition, Prentice Hall.

3. Burt, P. J. and E.H. Adelson. "The Laplacian Pyramid as a Compact Image Code." IEEE Trans. on Communications. April 1983. pp. 532 – 540.

4. Rafael C. Gonzalez and Richard E. Woods (2007), Digital Image Processing, 2nd Edition, Prentice Hall.

5. The Research Bulletin of Jordan ACM Vol 2

6. Charles Boncelet (2005)."Image Noise Models". in Alan C. Bovik. Handbook of Image and Video Processing.

7. J. C. Russ, *The Image Processing Handbook*, Boca Raton, FL: CRC Press, 1995.

8. A. C. Bovik, T. S. Huang, D. C. Munson, Jr., A generalization of median filtering using linear combinations of order statistics, *IEEE Trans. Acoust. Speech Signal Process.*

IT 039

Spatial Data Mining in Distributed DBMS

Name : Prof. Hidayatulla K. Pirjade
Institute: ASM'S Institute of Computer Studies.
Designation :Associate Professor
Email: hidayat.pirjade@gmail.com

ABSTRACT : Knowledge discovery in databases (or data mining) has become an important research area. Data mining is a nontrivial process to extract implicit, previously unknown, and potentially useful information from data in database systems. Geo-spatial data mining a subfield of data mining, is a process to discover interesting and potentially useful spatial patterns embedded in spatial databases. Efficient tools for extracting information from geo-spatial data sets are crucial to organizations which own, generate and manage large geo-spatial data sets. These organizations are spread across many domains including ecology and environmental management, public safety, transportation, public health, business logistics, travel, and tourism.

Challenges in spatial data mining arise from the following issues. First, classical data mining generally works with numbers and categories. In contrast, spatial data is more complex and includes extended objects such as points, lines, and polygons. Appropriate spatial modeling is required for analyzing and mining spatial datasets. Second, classical data mining usually deals with explicit inputs, whereas spatial predicates (e.g. overlap) and attributes (e.g. distance, spatial auto-correlation) are often implicit. Third, classical data mining treats each input as independent of other inputs, whereas spatial patterns often exhibit continuity and high spatial autocorrelation among nearby features. Fourth, modeling spatial context (e.g. autocorrelation) is a key challenge in classification problems that arise in geospatial domains. The current approach towards solving spatial data mining problems is to use classical data mining tools after "materializing" spatial relationships and assuming independence between different data points. However, this approach violates the key property of spatial data, which is spatial autocorrelation. Like temporal data, spatial data values are influenced by values in their immediate vicinity.

Ignoring spatial autocorrelation in the modeling process leads to results which are a poorfit and unreliable.

Keywords :*Heterogeneous Distributed Database, Collaborating Multiple Servers, Inefficiency of Data mining Techniques, Conceptual Middle ware standard Data Model, Distributed Database Systems ,Distributed Data Mining and Mining Multi-Agent Data,Mining effect.*

Introduction:

Data Mining is an analytic process designed to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment (i.e., the application of the model to new data in order to generate predictions).

A distributed database system consists of loosely coupled sites that share no physical component. Database systems that run on each site are independent of each other; transactions may access data at one or more sites. There are two types of distributed databases;

1. Homogeneous distributed database
2. Heterogeneous distributed database

If data is distributed but all server runs same DBMS software then it is called Homogeneous distributed database and if different sites run under the control of different DBMSs essentially autonomously and a connected somehow to enable access to data from multiple sites i.e. referred as Heterogeneous distributed database.

Finally as solution of the above mentioned problem is to be converting the multiple site data into the Single Standard Model this will be efficient in terms applying data mining techniques.

1. Heterogeneous Distributed Database

In this case different sites may use different schemas and software. So it leads to a major problem for query & transaction processing. Sites may not be aware of each other and may provide only limited facilities for cooperation in transaction processing.

2. Collaborating Multiple Servers

The client server architecture doesn't allow a single query to span multiple servers because of multiple sites having different DBMSs.

3. Inefficiency of Data mining Techniques on different DBMS models

All the sites are loosely coupled so that every site having different database management system which depends upon various DBMS models. While mining data through these DBMS models require distinct data mining techniques. So it leads to major problems like reliability, security, portability, performance, query & transaction processing.

4. Conceptual Middle ware standard Data Model

The conceptual middleware standard data model will allow to convert different data model into single portable format. So it may possible that to implement suitable data mining technique onto standard data model. This model should be more beneficial to new version database models and also avoids above mentioned problems.

Distributed Database Systems

A distributed database system contains a distributed database (DDB) and the software to manage the DDB. In a distributed database system, multiple logically related databases can be distributed over a computer network. Many machines can be used to solve one problem and data may be replicated at different sites, and hence improve performance, reliability, availability and extendibility. Applications of distributed database systems include multi-plant manufacturing, a chain of department stores, a bank with many branches, airlines, and military command and control.

Research issues in distributed database systems include distributed database design,query processing, concurrency control, fragment replication, and reliability. This research paper concentrates on the

optimization problem of data replicas. Data replication is an important research topic in distributed database systems . Such systems are built in a computer network with a certain topological structure. Multiple copies of data replicas are placed at different locations to achieve high data availability in the presence of possible network link failures. The best placement of fragments for a given number of data replicas in a computer network to minimize query response time and maximize throughput is a very difficult problem and has been studied extensively in the literature for various protocols. we presented a sufficient and necessary condition for the optimality of a placement of an odd number of data replicas in a ring network with a majority voting protocol.

Current work on spatial data mining has focused on predicting location problems. We have proposed PLUMS (Predicting Locations Using Map Similarity) which is need before conversion of objects in common model , a new framework for supervised spatial data mining problems. PLUMS searches the space of solutions using a map-similarity measure, which is more appropriate in the context of spatial data. We have shown that compared to state-of-the-art spatial statistics approaches such as the SAR (Spatial Autoregression Model) model ,PLUMS achieves comparable accuracy but at a fraction of the computational cost. Furthermore, PLUMS provides a general framework for specializing other data mining techniques for mining spatial data. We have also exploited different classification approaches for modeling spatial context (e.g. spatial autocorrelation) in the framework of spatial data mining. We compared and contrasted MRF (Markov Random Fields) and SAR using a common probabilistic framework. MRF is a popular model to incorporate spatial context into image segmentation and land-use classification problems. The SAR model, which is an extension of the classical regression model for incorporating spatial dependence, is popular for prediction and classification of spatial data in regional economics, natural resources, and ecological studies. Our study shows that the SAR model makes more restrictive assumptions about the distribution of features and class shapes (or decision boundaries) than MRF. We also observed an interesting relationship between classical models that do not consider spatial dependence and modern approaches that explicitly model spatial context. The relationship between SAR and MRF is analogous to the relationship between Logistic Regression and Bayesian Classifiers . We have provided theoretical results using a probabilistic framework as well as experimental results validating the comparison between SAR and MRF.

Distributed Data Mining and Mining Multi-Agent Data :

The problem of distributed data mining is very important in network problems. In a distributed environment (such as a sensor or IP network), one has distributed probes placed at strategic locations within the network. The problem here is to be able to correlate the data seen at the various probes, and discover patterns in the global data seen at all the different probes. There could be different models of distributed data mining here, but one could involve a NOC that collects data from the distributed sites, and another in which all sites are treated equally. The goal here obviously would be to minimize the amount of data shipped between the various sites — essentially, to reduce the communication overhead. In distributed mining, one problem is how to mine across multiple heterogeneous data sources: multi-database and multi-relational mining. Another important new area is *adversary data mining*. In a growing number of domains — email spam, intrusion detection/computer security, click spam, search engine spam, surveillance, fraud detection, shopbots, file sharing, etc. — data mining systems face adversaries that deliberately manipulate the data to sabotage them (e.g. make them produce false negatives). We need to develop systems that explicitly take this into account, by combining data mining with game theory.

The Mining effect:

Yes,data mining can have an effect. And I'm not thinking about the effect when you do the most exciting job in the world. I want to discuss the effect of applying data mining in an iterative way (call it the data mining bias if you prefer). Let me explain this topic with a concrete example.Think about market basket analysis (association rules mining). Imagine a case study that can happen at Amazon (to cite the most well known). We collect transactions made by customers. We build a model to suggest other books you may purchase. One month later, we run the same process again. However, the data collected was biased by the previous model. After several iterations, we may miss important associations if customers mainly buy what they are recommended.In another perspective, one may think that data mining (particularly in this case) limits the choices of the user.

Conclusion:

Data mining systems aim to discover patterns and extract useful information from facts recorded in databases. A widely adopted approach to this objective is to apply various machine learning algorithms to compute descriptive models of the available data. Here, we explore one of the main

challenges in this research paper, the development of techniques that scale up to large and possibly physically distributed databases. Distributed systems may need to deal with heterogenous platforms, with multiple databases and (possibly) different schemas, with the design and implementation of scalable and effective protocols for communicating among the data sites, and the selective and efficient use of the information that is gathered from other peer data sites. Data mining systems that must not be ignored, include, first, the ability to take advantage of newly acquired information that was not previously available when models were computed and combine it with existing models, and second, the flexibility to incorporate new machine learning and evaluate various proposed solutions(new model) through extensive empirical studies.

References:

- Database Management System- Henry Korth
- Advanced Database Management System- Ramkrishnan and Gherke
- Research Methodology, C.R. Kothari, New Age Publications,
- S. P. Gupta, *Statistical Methods*, Sultan Chand, New Delhi.
- W. Cheung, et al., "A Fast Distributed algorithm for Mining Association Rules," Proc. Parallel and Distributed Information Systems, IEEE CS Press, 1996,pp. 31-42;
- Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 2008-12-17.
- "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.
- Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". etrieved 2010-12-09.
- Ian H. Witten; Eibe Frank; Mark A. Hall (30 January 2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3 ed.). Elsevier. ISBN 978-0-12-374856-0. Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0471228524. OCLC 50055336.

IT 040

Enhance the privacy issues related to SQL Azure

Name: Mrs.Priya Joshi

Designation:-Assistant Professor

Institute:-ASm's ICS pimpri Pune

Email:-priya.r.n@gmail.com

Abstract:-

Privacy and Security are two of the key concerns most of us have, when thinking about storing our mission critical data in the cloud. There are people who appreciate the benefits storing data in a cloud based data store such as SQL Azure. There are also people who warn about possible dangers in storing important data in a cloud store.

SQL Azure provides relational database capability in the cloud. One of the features that is missing is a way to move databases up or down from your in-house SQL server. So how do you move a database schema and data up to the cloud? In this tip I will walk through a utility that will help make the migration much easier.

SQL Azure is Microsoft's relational database that is part of its Windows Azure Cloud Platform as a Service offering. While it includes most of the features of SQL Server 2008 it doesn't include any backup or restore capabilities that allow for hoisting schema and data from an on-premises database up to SQL Azure. The documentation refers to using SQL Server Management Studio's (SSMS) scripting capability for this task.

While SSMS has the ability to script both schema and data there are several problems with this approach:

- SSMS scripts all database features, but there are some features that SQL Azure doesn't support
- SSMS doesn't always get the order of objects correct
- SSMS scripts data as individual insert statements, which can be very slow

SQL Azure is based on SQL Server 2008, but the features that Azure supports are not identical to terrestrial SQL Server 2008. The Azure T-SQL is a subset of the full 2008 T-SQL. The analysis step examines scripts looking for features or syntax that doesn't work in SQL Azure. Some typical issues are:

- Tables without clustered indexes
- Cross database references

- use of SQLCLR or one of the built-in data types that depend on the CLR such as hierarchyid

Introduction:-

There are many definitions of cloud computing and some people think it's simply having a shared or dedicated machine at an ISP. Granted that machine is hosted in a data center that's more secure with backup power and high-speed internet connections. And it's certainly better than a box sitting in the network room down the hall, but that's not what we think about when we discuss cloud computing. That's just an off-site hosting of your computer.

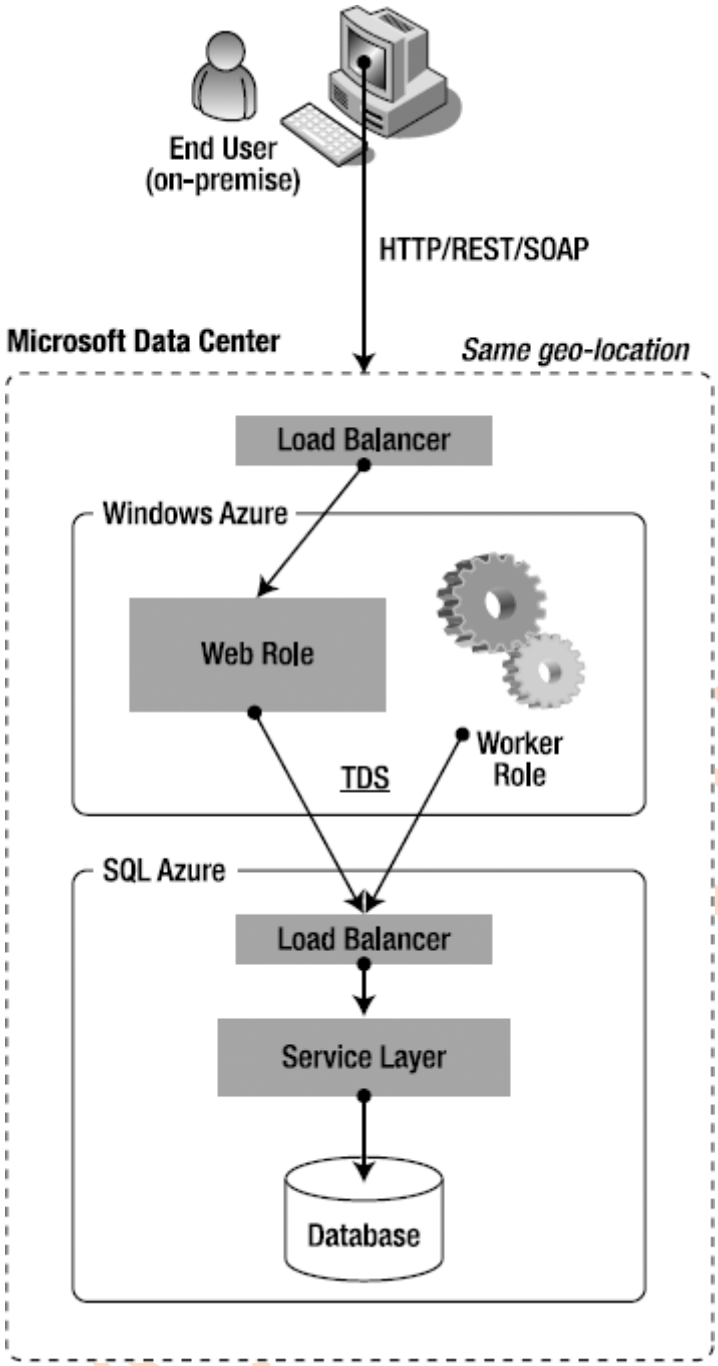
The clouds computing that excites us are the solutions like Microsoft Azure where the application resides on multiple machines automatically. This provides superior reliability and fault-tolerance compared to hosting in one server. It's designed to automatically adjust its resources as demand increases or decreases whether it's computing power, bandwidth or storage. Cloud providers like Azure also take advantage of edge networks to store/cache data closer to recipients to reduce congestion centrally.

These are truly enterprise and multi-national features that are nearly impossible to replicate on your own or with local ISP hosting companies. SQL Azure allows you to connect to the cloud database only using the TDS protocol with limited support, as described in the previous section. But because the TDS protocol is supported by most of the SQL Server client APIs, all the features supported by SQL Azure work with existing client APIs. You can use two common patterns to connect to SQL Azure databases: code near and code far.

1. Code-Near Connectivity

In code-near connectivity, your application is deployed in Windows Azure, which uses SQL Azure. You geo-locate both of them in the same data center by configuring the geo-location features of Windows Azure and SQL Azure. Figure 1 illustrates applications with code-near connectivity to a SQL Azure database.

Figure 1. Code-near connectivity to SQL Azure



In typical code-near architecture, the data access application is located in the same data center as the SQL Azure database. The end users or on-premises applications access the web interface are exposed via a Windows Azure web role. This web role may be hosting an ASP.NET application for end users or a web service for on-premises applications.

The advantages of the code-near approach are as follows:

- Business logic is located closer to the database.

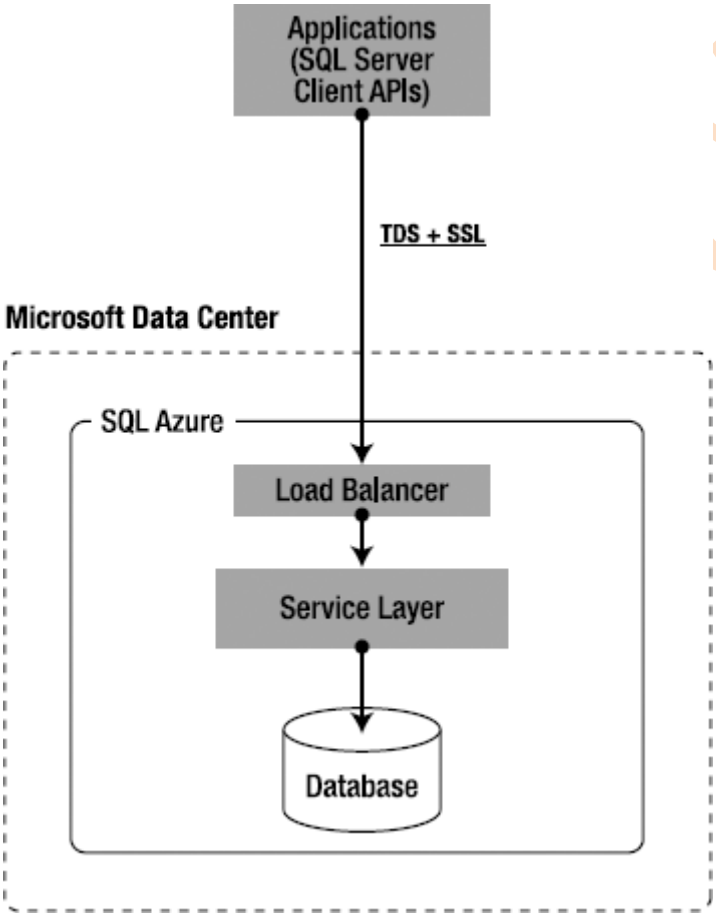
- You can expose open standards–based interfaces like HTTP, REST, SOAP, and so on to your application data.
- Client applications don't have to depend on the SQL Server client API.

The disadvantage of this approach is the performance impacts your application experiences if you're using Windows Azure as a middle tier to access the database.

2. Code-Far Connectivity

In code-far connectivity, your application is typically deployed on-premises or in a different data center than SQL Azure. In this pattern, the client application makes a SQL query using the TDS protocol over the Internet to the SQL Azure database. Figure 2 illustrates applications with code-far connectivity to a SQL Azure database.

Figure 2. Code-far connectivity to SQL Azure



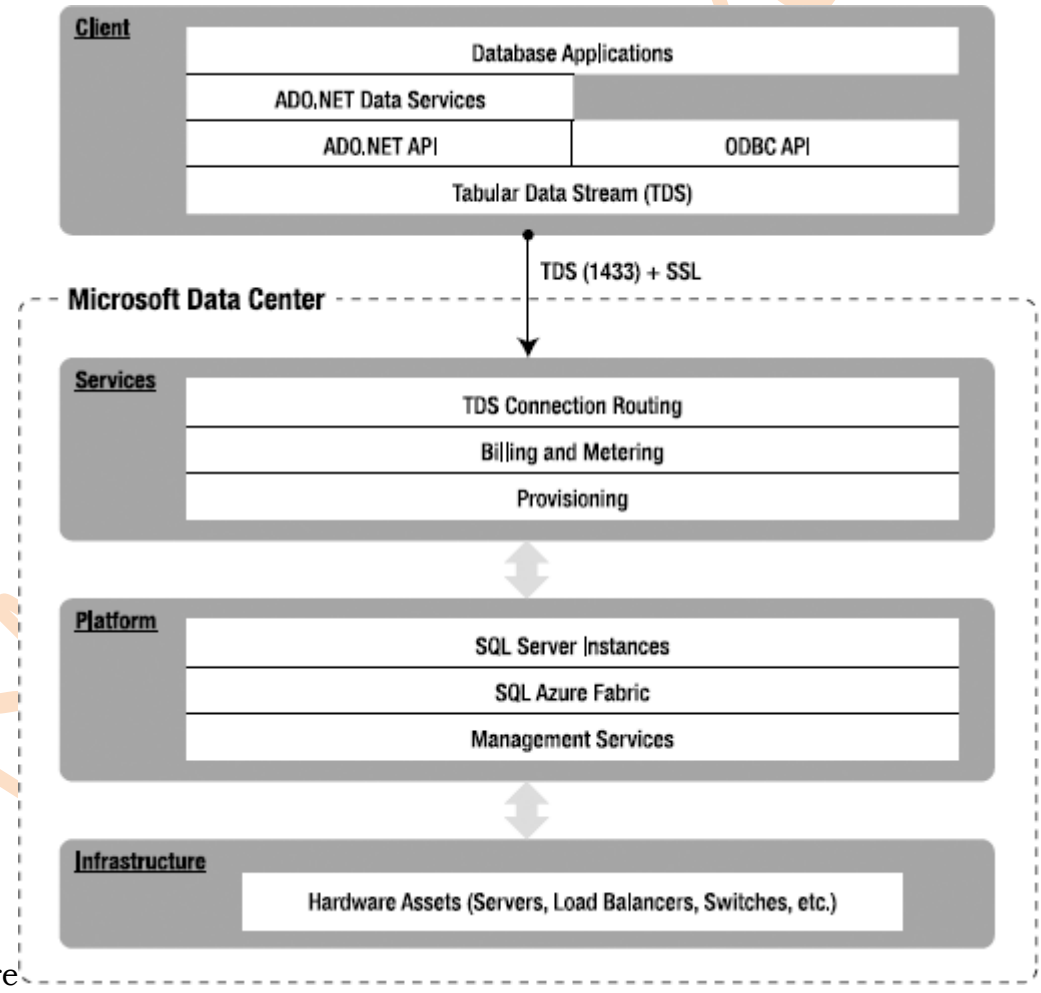
The biggest advantage of the code-far approach is the performance benefit your application can experience because of direct connectivity to the database in the cloud. The biggest disadvantage is that all the client applications must use the TDS protocol to access the database. Therefore, the data access clients must use SQL Server–supported client APIs like ADO.NET, ODBC, and

so on, reducing data-access possibilities from APIs or platforms that don't support the TDS protocol.

SQL Azure Architecture :-

sql azure is a scalable and highly available database utility service in the cloud. like all other windows azure services, it runs in microsoft data centers around the world. the data center infrastructure provides the sql azure service with load balancing, failover and replication capabilities. figure 1 illustrates the high-level sql azure architecture.

Figure 1. SQL Azure Architecture



As shown in Figure 1, the SQL Azure service is composed of four layers: infrastructure, platform, services, and client. All the layers except the client layer run inside a Microsoft data center.

1. Infrastructure Layer

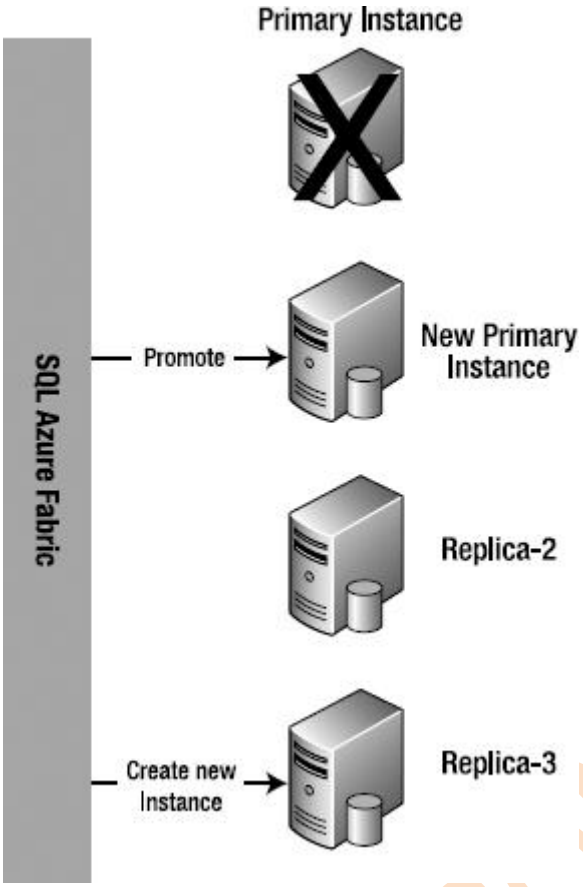
The infrastructure layer is the supporting layer providing administration of hardware and operating systems required by the services layer. This is the core data center layer that is shared across multiple services in a data center.

2. Platform Layer

The platform layer consists of the SQL Server instances and the SQL Azure fabric, and Management services. The SQL Server instances represent the deployed databases, their replicas, and the operating system instances that host the SQL Server instances. The SQL Azure fabric is the underlying framework that automates the deployment, replication, failover, and load balancing of the database servers.

The SQL Azure fabric is responsible for creating three replicas of your database instance and provides automatic failover capabilities to these instances. As shown in Figure 2, if the primary instance of your database experiences a failure, the SQL Azure fabric designates one of the replicas as the primary instance and automatically routes all the communications to the new primary instance. In an effort to maintain three replicas at all times, SQL Azure also creates a new replica of the database.

Figure 2. SQL Azure database replicas



The Management services are responsible for maintaining the health, upgrades, consistency, and provisioning of the hardware and software to support the SQL Azure fabric.

3. Services Layer

The services layer comprises external (customer) facing machines and performs as a gateway to the platform layer. It exposes the tabular data stream (TDS), billing, metering, and account provisioning services to customers.TDS is the native Microsoft SQL Server protocol that database clients can use to interact with a SQL Server database.The services layer exposes the TDS protocol on port 1433 over Secure Sockets Layer (SSL). The services layer is also responsible for routing connections to the primary database instance in the platform layer. This layer maintains runtime information about your database replicas and routes the TDS coming from client applications to the appropriate primary instance. The services layer is also responsible for provisioning your database when you create a database in SQL Azure. The provisioning of databases involves communicating with the SQL Azure fabric in the platform layer to provision appropriate replicas of the database.

The billing and metering service is responsible for monitoring the runtime usage of your database for billing purposes. The billing and metering service tracks the usage of databases at the account level.

4. Client Layer

The client layer is the only layer that runs outside of the Microsoft data center. The client layer doesn't include any SQL Azure–specific components; instead, it uses all the existing features of SQL Server client components like ADO.NET, ODBC, Visual Studio.NET, SQL Server Management Studio, ADO.NET Data Services, and so on. The client API initiates a TDS connection to SQL Azure on port 1433, which is routed by the services layer to the platform layer to the appropriate database instance.

Windows Azure is an open cloud platform that enables you to quickly build, deploy and manage applications across a global network of Microsoft-managed datacenters. You can build applications using any language, tool or framework.

Benefits of SQL Azure Database

No physical administration required: spend your time designing, optimizing and enabling solutions

Because SQL Azure is a managed service, you do not have to install, set up, patch or manage hardware or software. There is no need to create or manage your own virtual machines or roll your own high availability. Every SQL Azure database has built-in high-availability, failover, and redundancy.

Scale On-Demand to meet your business needs

SQL Azure gives you the flexibility to scale out depending on your needs while paying only for what you use. Rely on business-ready SLAs and global datacenters to offer highly available services to your users or to build out your own multi-tenant offerings.

Managed Service

Data Sync is a fully managed cloud-based service. There is no need to write complex database logic to synchronize and transfer data between databases. Instead, simply use the point-and-click portal to quickly configure and schedule synchronization.

On-Premises and Cloud

Enables multiple synchronization scenarios spanning both cloud and on-premises databases. Now it is easy to enable one-way as well as bi-directional data movement across SQL Azure and on-premises SQL Server databases.

Cloud-to-Cloud Synchronization

Data can be shared between multiple databases, irrespective of whether the databases are in the same data center or span multiple geographic regions.

Control Synchronization: Specify exactly what tables and columns to synchronize, setup filters to sync only a subset of rows, set your conflict resolution policy for two-way sync, and specify how frequently data should be synchronized.

Methodologies:-

SQL Azure is Microsoft's relational database that is part of its Windows Azure Cloud Platform as a Service offering. While it includes most of the features of SQL Server 2008 it doesn't include any backup or restore capabilities that allow for hoisting schema and data from an on-premises database up to SQL Azure. The documentation refers to using SQL Server Management Studio's (SSMS) scripting capability for this task.

While SSMS has the ability to script both schema and data there are several problems with this approach:

- SSMS scripts all database features, but there are some features that SQL Azure doesn't support
- SSMS doesn't always get the order of objects correct
- SSMS scripts data as individual insert statements, which can be very slow

Recognizing that the SSMS approach doesn't work very well, two programmers, George Huey and Wade Wegner created the SQL Azure Migration (SAMW) wizard and posted it on CodePlex. You'll find the project on the SQL Azure Migration Wizard page where you can download the program including the source code, engage in discussions and even post a patch if you're interested in contributing to the project.

You'll need the SQL Server R2_client tools (November CTP or later) and the .Net Framework on the machine where you unzip the download file. There's no install program, at least not yet, just the SQLAzureMW.exe program and configuration files.

SAMW is a wizard style Windows Forms program. When the SAMW is started the first task is to select a process to apply. The available options, shown below, allow you to analyze and move schema and data between databases on the local network and SQL Azure databases. For the purposes of this article we'll follow the Analyze and Migrate Wizard shown selected here.

One of the features that make SAMW much more than a scripting tool is the optional analysis step. SQL Azure is based on SQL Server 2008, but the features that Azure supports are not identical to terrestrial SQL Server 2008. The Azure T-SQL is a subset of the full 2008 T-SQL. The analysis step

examines scripts looking for features or syntax that doesn't work in SQL Azure. Some typical issues are:

- Tables without clustered indexes
- Cross database references
- use of SQLCLR or one of the built-in data types that depend on the CLR such as hierarchyid

SAMW will find these issues and many more by examining the T-SQL using regular expressions and can in addition, suggest or make changes. In particular, it does a good job of adding clustered indexes to tables that don't have them.

As you can see from the choices above there are three places that the T-SQL used for the analysis can come from:

- The Database
- A file of T-SQL code
- A SQL Profiler Trace file

Any of these sources is analyzed by applying a set of regular expressions that live in SAMW's NotSupportedByAzureFile.Config file. The top of the file, formatted for visibility, is shown here with its first rule that removes "NOT FOR REPLICATION" clauses.

The NotSupportedbyAzureFile can be customized to make additional changes, add additional messages, or to relax restrictions as SQL Azure evolves.

Once we've chosen the "Analyze and Migrate\SQL Database" wizard and press "Next >" the standard select a server dialog appears and a source database must be selected

The screen shows picking the old pubs database. It's still possible to download the pubs and Northwind databases from the Microsoft download site. Once a database is chosen it's possible to select all objects or a subset of the database objects to analyze and move. In this case I've chosen to move all objects

A summary screen is displayed and then SAMW goes to work scripting the selected database objects and analyzing them for issues based on the NotSupportedByAzureFile and then BCPing the data out to temporary files.

The pubs database is pretty simple and the only problems that it found were missing clustered indexes on two tables: dbo.discounts and dbo.roysched. SAMW picks a column to cluster on arbitrarily and moves forward to the BCP step. This screen shot illustrates how the report shows what it’s done:

After the analysis is done the user must proceed to connect to SQL Azure as shown below. Notice the server name is followed by ".database.windows.net" and the server name, proceeded by an @ sign, must go in the User name field after the actual user name. This extra server name is necessary because some connection methods need the server name in the UserID field of the connection string. Parts of the server name and user name below have been obscured for privacy reasons.

The final step is to create the target database for migration and push the trigger. The upload process for pubs isn't very long because the database is so small.

The upload rates are much slower that come to expect from BCP. Even going to a slow server I'd expect ten times the number of rows per second. The slower performance may have a lot to do with the transmission over the internet, but it also may be due, in part, to two SQL Azure features:

- The data is written two three copies of the database
- Clustered Indexes are required instead of heaps, which are best for BCP

All in all the SQL Azure Migration Wizard is a necessary tool for anyone working with SQL Azure in the short term. When additional features are added to SQL Azure, such as SQL Server backup and restore, SAMW may not be ne

Conclusion

Client-plus-cloud computing offers enhanced choice, flexibility, operational efficiency and cost savings for businesses and consumers. To take full advantage of these benefits, users must be given reliable assurances regarding the privacy and security of their online data. In addition, a number of regulatory, jurisdictional, and public policy issues remain to be solved in order for online computing to thrive.

The components anchor our commitment to maintaining the highest standards of privacy and security in our online services and to partnering with other industry leaders, governments, and consumer organizations to develop globally consistent privacy frameworks that enable the expansion of the economic and social value of cloud-based computing.

References:-

1. **http://www.windowsazure.com**
2. **http://www.microsoft.com/windowsazure/**
3. **http://www.mssqltips.com/sqlservertip/1989/sql-azure-migration-wizard/**

IT 041

"FREE/OPEN SOURCE SOFTWARE:

OPPORTUNITIES AND CHALLENGES FOR MODELING AND SIMULATING "

Mrs. Ashwini R. Bhirud

Assistant Professor, ASM's IBMR Institute, Pune

ashwini.rbhirud@gmail.com

ABSTRACT:

This article introduces a special issue of Software Process-Improvement and Practice focusing on processes found in open source software development (OSSD) projects. We are interested in process discovery in large, globally distributed organizations which currently engages over twenty thousand developers distributed over several continents working collaboratively, sometimes across several stages of the software lifecycle in parallel. This presents a challenge for those who want to join the community and participate in, as well as for those who want to understand these processes. It presents, analyzes, and compares data collected from different F/OSSD projects, including an in-depth case study, to help develop such an understanding. The goal of this chapter is to determine the circumstances and conditions when F/OSSD represents a viable alternative to SE for the development of complex software systems. These concepts include discovering and understanding: what affects the cost, quality, and productivity of F/OSS development; F/OSS evolution dynamics; why people participate, join, and contribute to F/OSS projects; continuous F/OSS processes; multiproject dependencies. We identify how these concepts stimulate process modeling and simulation challenges and opportunities.

Keywords: Open source software, software process modeling, software process simulation.

INTRODUCTION:

The free/open source (F/OS) software development model gives organizations a new option for acquiring and implementing systems, as well as new opportunities for participating in F/OS projects. What does this mean in practice for libraries and information centers, which have specialized requirements and make extensive use of technology to provide services to their users? This paper begins with a description of F/OS concepts, followed by a discussion of the benefits and issues associated with this approach. The paper concludes with a discussion of factors associated with the successful implementation of F/OS software.

Free/open source concepts

If asked, most computer-literate people would say that free/open source software includes the original code for the program, whatever language it is written in (C, C++, Perl, PHP, etc.) and that it also may be modified for local use and then subsequently redistributed for "free". However, the official definitions of "free software" and "open source" cover other aspects of software use and distribution that are important in understanding how this type of software differs from commercial or proprietary software.

The Free Software Definition

The Free Software Foundation (FSF) maintains identifies four aspects of freedom, from the software users' perspective:

- The freedom to run the program, for any purpose (freedom 0)
- The freedom to study how the program works, and adapt it to local needs (freedom 1)
- The freedom to redistribute copies so others can benefit from the software (freedom 2)
- The freedom to improve the program, and release the improved version to the public, so that the community can benefit (freedom 3).

In order for software to qualify as "free" under freedoms 1 and 3, users must have access to the source code.

The Open Source Definition

The Open Source Initiative (OSI) maintains the Open Source Definition (OSD), which has 9 clauses:

1. Software must be able to be freely distributed, without requiring a royalty or fee for sale.

2. The source code for the program must be available, and, if not included in a distribution, must be easily available (for example, downloadable from a web page) in a form which allows a programmer to modify the program.

3. Modifications and derived works must be allowed, and these must be able to be redistributed under the same terms as the original software.

4. The integrity of the original source code must be able to be maintained, either by requiring modifications to be distributed as "patch files", or by requiring modified versions to have a different name or version number.

5. There must be no discrimination against persons or groups.

6. There must be no discrimination against any field of Endeavour.

7. The license must apply to anyone receiving a copy of the program, without requiring them to agree to another license.

8. The license must not be specific to a particular product or distribution.

9. The license must not apply to other software distributed along with the licensed program(s) .

Free software vs open source software

Though these two definitions represent very different philosophical positions, with the OSI being more pragmatic and the FSF less compromising, in practice they have a broad overlap. A wide range of F/OS software is available, including the Linux operating system, the Apache web server, and the Send mail utility that is used by many mail servers. Open source programming languages include Perl and PHP, and open source database management software such as MySQL and PostgreSQL is also widely used.

One particularly important point about both definitions is that they are about terms and conditions for distributing software, and say nothing about the methods and processes used to develop it—but because source code is available and can be modified by anyone who has the skill and interest, F/OS software can evolve differently from commercial packages. One effect of the use of F/OS software licensing is the development of communities of developers and users around specific F/OS projects.

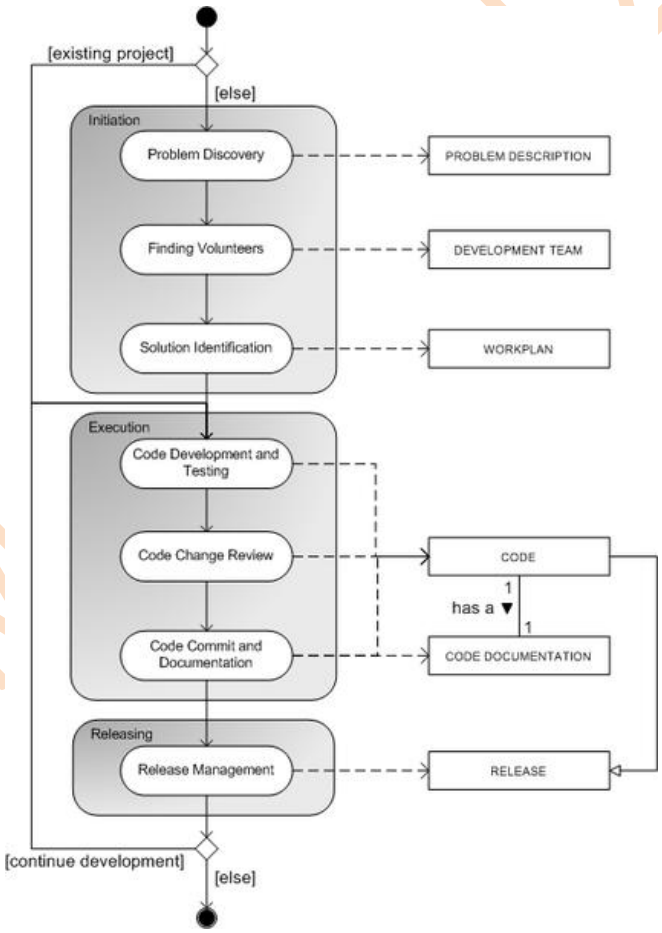
Free/open source software development

Open source software development is the process by which open source software (or similar software whose source code is publicly available) is developed. These are software products "available with its source code and under an open source license to study, change, and improve its design". Examples of popular open source software products are Mozilla Firefox, Google Chromium, Android and the OpenOffice.org Suite. In the past, the open source software development method has been very unstructured,

because no clear development tools, phases, etc., had been defined like with development methods such as Dynamic Systems Development Method. Instead, every project had its own phases. However, more recently there has been much better progress, coordination, and communication within the open source community.

Open source software development can be divided into several phases. A diagram displaying the process-data structure of open source software development is shown on the right. In this picture, the phases of open source software development are displayed, along with the corresponding data elements. This diagram is made using the meta-modeling and meta-process modeling techniques.

The process starts with a choice between the adopting of an existing project, or the starting of a new project. If a new project is started, the process goes to the Initiation phase. If an existing project is adopted, the process goes directly to the Execution phase.



Dig. Process-data diagram of Open source software development

For the purposes of this article, focus is directed at F/OSSD processes, rather than to software licenses and social movements with free or open source software, though each may impinge on F/OSSD processes.

F/OSSD is mostly not about software engineering, at least not as SE is portrayed in modern SE textbooks. F/OSSD is not SE done poorly. It is instead a different approach to the development of software systems where much of the development activity is openly visible, development artifacts are publicly available over the Web, and generally there is no formal project management regime, budget or schedule.

F/OSSD is also oriented towards the joint development of a community of developers and users concomitant with the software system of interest. There is also widespread recognition that F/OSSD projects can produce high quality and sustainable software systems that can be used by thousands to millions of end-users. Thus, it is reasonable to assume that F/OSSD processes are not necessarily of the same type, kind, or form found in SE projects that follow the processes described in modern SE textbooks.

Success factors for adopting F/OS software

In many ways, adopting open source software requires the same type of evaluation as purchasing commercial software, but with some modification. One analogy that is frequently used to describe using F/OS software is that it is like getting a "free" puppy or kitten. While the initial cost is zero, supporting the software over its lifetime may involve considerable costs, and organizations may find it difficult to estimate these costs when the software is initially adopted. To some organizations, the advantages of having more control over the software's future development will outweigh the perceived risks of using it, but others may come to the opposite conclusion.

Libraries thinking of adopting a F/OS software application need to consider the following factors.

Organizational culture

What support is there within your library and its parent organization for the F/OS philosophy, and at what level? Is the organization willing to take some responsibility for the future development of the software?

Technical infrastructure

What operating system(s) does the software run on, and does your organization support these? Is the programming language (Perl, PHP, etc.) installed and available? Is the database application (e.g. MySQL or PostgreSQL) available? What additional technologies does the software require

(e.g. Apache web server)? Is there adequate documentation about the requirements and installation procedures? Can you try the software without affecting your current systems?

Staff skills

Are staffs familiar with the programming language and operating environment? Will the organization support any staff training and development needs to support the F/OS application? If not, do you have access to an external organization that can provide the necessary support?

Software functionality

How mature is the software? How stable is it? How well does its functionality match your requirements? If it will need changes to meet your specific needs, do your staff have the skills and time to make these changes? Is there a process to have these changes incorporated into the package for future releases? What is the perceived risk of using the software? Using F/OS software to provide core functionality, such as a library management system, might be seen as too risky, while using F/OS software that provides stand-alone functionality that complements current systems might be acceptable.

Project community

How active is the project's community, and what resources are available to support new users of the software? What would you be able to contribute back to the community? Where is community resources such as email archives and documentation hosted?

Process Modeling and Simulation

There are variety of new challenges for modeling and simulating F/OSSD processes. As studies of F/OSSD processes may include analytical tools, techniques, or strategies from multiple, so should the process modeling and simulation methods acknowledge and employ these discipline-specific capabilities.

- New kinds and types of software processes have been observed and studied in F/OSSD projects and exponential evolutionary growth of the most successful F/OSSD project's source code base. How are these processes most effectively modeled?

- Basic software development processes associated with requirements development and software design seem to be primarily dependent of the use of software in formalisms. As such, what is the form of software processes that can be modeled to account for the production and consumption of F/OSSD in

formalisms through conversational or computer-mediated communication systems?

•Most F/OSSD projects do not succeed in achieving a critical mass of core developers, but those that do often develop their software systems in a continuously emerging manner. Thus, yesterday's system and functionality is not tomorrow's system and functionality. As such, what is needed to model software processes that are continuously evolving and adapting to new circumstances, platforms and applications.

The diversity of this set of enumerated software process modeling challenges points to the richness of the field of software process research targeted to modeling F/OSSD processes. Simulating F/OSSD processes also poses a complementary diversity of challenges.

F/OSSD process modeling and simulation research:

It appears that the most likely research focus in F/OSSD process modeling and simulation will examine one or more of the following:

• Exploratory studies of new F/OSSD process types;

• Modeling and simulation of F/OSSD processes;

• Multi-disciplinary modeling and simulation of F/OSSD processes

• Modeling and simulation of continuous F/OSSD processes.

Strategies that support of the F/OSSD process modeling and simulation will examine either:

• *Single process within a single project,* is important when examining high profile F/OSSD projects, where the selected process may be unique or untypical of other F/OSSD projects.

• *Multiple processes within a single project,* where focuses attention to a high-profile F/OSSD project in order to account for some overall development phenomena.

• *Single process found in multiple projects*, where emphasis is on understanding the form and variations of the selected F/OSSD process through comparative analysis across F/OSSD projects.

• *Multiple processes found in multiple projects*, where emphasis is on understanding the form and variation of overall F/OSS development or evolution process, across projects over time. Sub-samples may further focus attention to F/OSSD processes within multiple projects of a common type (e.g., Internet infrastructure or networked computer games), and finally

multiple projects across multiple project types (infrastructure, games, and science).

- *Population studies* are focuses on studies that seek to characterize the overall population or universe of F/OSSD projects, as perhaps might be associated with a specific F/OSS Web portal like SourceForge.org.

Data collection methodologies in support of F/OSSD process modeling and simulation include:

- *Ethnographic and qualitative field studies*, especially when emphasizing social, cultural, or socio-technical processes within a single project.
- *Case studies and comparative case studies*, when focusing on in-depth comparisons of processes of the same type in different F/OSSD projects, or more comprehensive studies that examine multiple types of processes across multiple types of F/OSSD projects.
- *Surveys, questionnaires, or online polls* are well suited when seeking to ascertain processes that are shaped by participant's perceptions, beliefs, or opinions, especially when large samples of participants are available.
- *convergence methods* seek to build on the use of many of the preceding data collection methodologies, so as to be able to characterize, model, and simulate F/OSSD process from multiple perspectives, supported by multiple kinds of process data.

Every empirical study requires or benefits from an explicit strategy for assuring the quality of the models and simulations produced. As before, a variety of choices are available, though they generally depend on choices made for the preceding framework components. The strategies seen in the surveyed studies cover the following range of assurance alternatives:

- *Packaging and fit* is the baseline form of assurance that addresses how the analytical variables were identified and composed, which determines whether the analysis presented makes sense, is coherent, fits the data to the model, and rules out other alternative explanations that would refute the model.
- *Reliability and construct validity* are often used to explain variance measures that result from a factor analysis of quantitative data. Such assurance is focused on quantitative process data.
- *External validity and traceability* focuses attention to whether the participants engaged in performing the process can make sense, can trace process fragments back to their source data, and be satisfied that the analysis offers them something of value, such as process improvement options.
- *Cross-comparative grounded theory* assures the resulting process model is based on data arising from comparative ethnographic methods. The process

model is composed, compared, cross-checked, and presented incrementally so as to provide a rich account of the process and data sources. New data will not refute the model, but instead may realize incremental model improvement and refinement.

- *Cross-domain theories* provide multi-discipline analytical methods and theoretical perspectives that collectively serve to explain the modeled process was constructed, what it explains, and what multi-disciplinary associations it makes.

The components of a framework that accounts for how F/OSSD process modeling and simulation studies may be organized or structured. Thus, it offers a huge selection of research opportunities and challenges that require further study and contribution.

CONCLUSIONS:

The purpose of this research paper was firstly to identify how the open source phenomenon can alleviate system development challenges faced by organizations. OSSD combines features found in traditional software processes with other features in a unique way that can potentially produce high-quality software, faster and cheaper within the rapidly changing Internet environment. It provides potential benefits and opportunities to the system development process. The most obvious benefit is the reduced cost that open source provides. In addition, open source offers new business models that can be exploited, particularly within developing economies, and has an added benefit of skills development.

F/OSSD offer new types and new kinds of processes to model and simulate. Similarly, understanding how F/OSSD processes are similar to or different from SE processes is an area ripe for further research and comparative study. Many new research opportunities exist in the empirical examination, modeling, and simulation of F/OSSD processes.

However, such data has often been scarce, costly to acquire, and is often not available for sharing or independent re-analysis for reasons including confidentiality or nondisclosure agreements. F/OSSD projects and project repositories contain process data and product artifacts that can be collected, analyzed, shared, and be re-analyzed in an free and open source manner. F/OSS thus poses the opportunity to favorably alter the costs and constraints of accessing, analyzing, and sharing software process and product data, metrics, and data collection instruments. F/OSSD is thus poised to alter the calculus of empirical software engineering, and software process modeling and simulation research is an arena that can take advantage of such a historically new opportunity.

Through a survey of F/OSSD projects and other analyses presented in this article, it should be clear there are an exciting variety and diversity of opportunities for new software process modeling and simulation research. Thus, you are encouraged to consider how your efforts to research or apply software process modeling and simulation concepts, techniques, or tools can be advanced through studies that examine processes found in F/OSSD projects.

REFERENCES:

1. http://www.ics.uci.edu/~wscacchi/Presentations/ProSim04/Keynote-Scacchi.pdf

2. http://en.wikipedia.org/wiki/Open_source_software_development

3. http://quintagroup.com/cms/open-source

4. http://www.isr.uci.edu/research-open-source.html

5. http://www.ics.uci.edu/~wscacchi/Papers/New/ProSim04-Scacchi.pdf

6. http://www.mendeley.com/research/understanding-open-source-software-evolution/

IT 042

Green IT – With Emerging Approaches

Prof. Asheesh Dixit[1] Santosh. B. Potadar[2] Yogeshkumar. V. Gaikwad[3]

Director Assistant Professor Assistant Professor

ICS, Pimpari ICS, Pimpari ICS, Pimpari

ABSTRACT

IT has become part and parcel of business processes across various industries and global pursue of going "Green" will remain unachieved unless IT enables efficient use of power and energy. A number of practices can be applied to achieve energy efficiency such as improvement of applications' algorithms to optimally utilize the processor by reducing the proceeding cycle(s) which can facilitate the minimal use of power by switching of idle nodes. Similarly major approach could be to have energy efficient hardware by the use of ACPI Open Industry standards Interface and Dynamic voltage and frequency scaling (DVFS). A latest computing paradigm which can play a significant role towards Green IT is the "Cloud Environment" which uses the concept of virtualization of computing resources. Data center requires great effort to maintain consistent power supply for cooling and infrastructure maintenance.

KEYWORDS

Green IT, Algorithmic Efficiency, ACPI, DVFS, Cloud Computing, Data center

1. INTRODUCTION

Green IT also strives to achieve economic viability and improved system performance and use. While abiding by our social & ethical responsibilities. IT has become part and parcel of business processes across various industries

and global pursue of going "green" will remain unachieved unless IT enables efficient use of power and energy.

A number of practices can be applied to achieve energy efficiency such as improvement of applications' algorithms to optimally utilize the processor by reducing the proceeding cycle(s) which can facilitate the minimal use of power by switching of idle nodes. Similarly major approach could be to have energy efficient hardware by the use of ACPI Open Industry standards Interface. Another very popular approach can be that of dynamic voltage and frequency scaling (DVFS).

A latest computing paradigm which can play a significant role towards Green IT is the "cloud environment" which uses the concept of virtualization of computing resources. Data center requires great effort to maintain consistent power supply for cooling and infrastructure maintenance. The different scenarios are enlightened here in correspondence of data center issues such as security, power & energy consumption. Regulation on building of data centers better technologies to do so.

Virtualization technology which is the essence of Cloud computing allows one to create several virtual machines (VMs) on a physical server and therefore, reduces amount of hardware in use and improves utilization of resources. Traditionally, an organization purchases its own computing resources and deals with the maintenance and upgrade of the outdated hardware, resulting in additional expenses. Cloud computing model encourages an organization to outsource their computation needs to the Cloud, thereby eliminating the need to maintain own computing infrastructure which in turn naturally leads to energy-efficiency.

The rest of this paper organized as follows section 2, describes how algorithmic efficiency helps in efficient power consumption. In section 3, it includes power management through two techniques namely, DVFS and ACPI. Section 4 gives how latest technologies cloud computing and virtualization performs optimal utilization of virtual machines, servers which achieve efficient power consumption.

The last section 5 is related to storage data centers which require huge infrastructure, power & energy cost. It also explains solar power data centers can be used as solution on power usage.

2. ALGORITHMIC EFFICIENCY

The efficiency of algorithms has an impact on the amount of computer resources required for any given computing function and there are many efficiency trade-offs in writing programs.

Design such kind of algorithms which has efficient time complexity. The efficiency of algorithm is counted by considering time and space complexity. Release computer resources as early as possible by applying quickly executable algorithms whose time complexity is less. As soon as computer resources gets freed, its power turned to off through self regulatory mechanism.

The major focus is to maintain time complexity which has to be reducing up to possible extent. In short an algorithm which has exponential complexity optimizes it up to cubic, quadratic, linear or logarithmic.

O(2n) ---> O(n3) ---> O(n2) ---> O(n log n) ---> O(n) ---> O(log n)

Consider small demonstration example of calculating GCD of two numbers. Finding GCD of two numbers is possible by using traditional method of factorization. But this method takes more time for execution if higher digit number appears. Execution time for same problem can be reduced by implementing [11] Euclid method. This method requires minimum iterations which is same as count of digit of number.

Algorithm for GCD of two positive numbers (Euclid Method)

1: **START**

2: **Read** two positive numbers a and b

3: i← a and j← b

4: **for** i from a **to** not equal to 0 and j from b **to** not equal to 0

5: **if** (a > b) **then** a = a%b **otherwise** b = b % a

6: **end if**

7: **end for**

8: **if** a is equal to 0 **then Display** b as GCD

9: **Otherwise Display** a as GCD

10: **STOP**

Generally above algorithm requires minimum iteration as compared to traditional method. Euclid Method [12] has O(n) complexity if both numbers having n digits. Complexity varies as two numbers inputted with different digits. Finally aim is to develop efficient algorithms by using sharp logic and efficient data structures.

While considering algorithm related to database application, implementation of disconnected database architecture should be useful to develop faster database algorithms. In case of database application it is also possible to group co-related queries into one unit and execute it generally at the same time or dynamically as needed.

3. POWER MANAGEMENT

3.1 DVFS (Dynamic Voltage and frequency Scaling)

Dynamic voltage frequency scaling (DVFS) reduces operating voltage [3] to the minimum level needed to support only the minimum operating frequency required by applications at any given moment. The i.MX31 and i.MX31L bring the concept to new heights by introducing an automatic hardware mechanism for DFVS control that requires minimum CPU resources and software complexity in the operating system and drivers.

The power reduction is an important design issue for energy constrained applications. Lowering the supply voltage is one of the attractive approaches to save power of the variable workload system. Furthermore, dynamical scaling the clock frequency and the supply voltage achieves extremely efficient energy saving. Steady increase of the functional complexity and integration

level of SoC (solid circuit) separates the chip into multiple the power domains that have different performance requirement. And each power domain operates for different workload with different supply voltage and clock frequency.

3.2 ACPI (Advanced configuration and power interface)

ACPI is an open industry standard [2] allows an OS to directly control the power saving aspects of its underlying hardware. This allows system to automatically turn off components such as monitors and hard drives after set periods of inactivity. In addition system may hibernate, where most components including the CPU and the system RAM are turned off. Some programs allows the user to manual adjust the voltages supplied to CPU which reduces both the amount of heat produced and electricity consumed this process is called undervolting. Some CPU can automatically under volt the processor depends on workload.

Power management can also be regulated by using upcoming concept of autonomic computing which provides self-regulatory mechanism.

4. CLOUD COMPUTING AND VIRTUALIZATION

Cloud Computing [6] is the latest paradigm in IT sector which provides everything as service to users. Many cloud providers are available to provide different types of services such as infrastructure as a service (IaaS), Platform as Service (PaaS), and Software as service (SaaS) to users. Now days Green IT or Computing plays vital role with cloud providers because they want to maintain huge environment for user. The cloud environment requires huge amount of power and energy. Hence to reduce operational cost regarding to power is the great challenge in front of cloud providers. While reducing power consumption in cloud environment, providers has also think about to meet [6] SLA without affecting performance and deadline constraint. Hence it is necessary that to consider two approaches-

4.1 Provision of virtual machines in cloud data centers using DVFS

The major way is to reduce power consumption in providing virtual machines [5] in cloud environment on the basis of DVFS scheme. It includes mapping of virtual machines to user for execution of their applications with various DVFS schemes [6]-

Lowest DVFS for Virtual Machine provisioning

These adjusts the processors speed to the lowest at which virtual machine meet their task deadline i.e. each virtual machine executes its service at the required MIPS rate.

Advanced DVFS for Virtual Machine provisioning

In order to overcome low service accepantance rate of lowest DVFS scheme this scheme over scales up to δ% of the required MIPS rate for current virtual machines. Thus it operates the processor δ% faster in order to increase the possibility of accepting incoming application requests.

Adaptive DVFS for Virtual Machine provisioning

When deadlines of applications known in advance then it is possible to allocate virtual machines on the basis of optimality principle. Find the processor which has highest speed of execution for requested task than assign it to that virtual machine.

4.2 To allocate minimum resources i.e. virtual machines dynamically for the execution of applications of users

The recently emerging cloud computing paradigm leverages virtualization of computing resources and allows the achievement of more efficient allocation of the workload in terms of higher resource utilization and decreased power consumption. Before assigning virtual machines to user check workload of each virtual machine. If workload founds to be null then turned off that virtual machine automatically in cloud environment. Virtual machines which are heavily overloaded then perform work load balancing and turned on other virtual machines which was previously turned to off because of null workload.

Another major aspect is to utilize resource allocation framework models and by applying algorithm which dynamically allocates virtual machines for the

execution of application as per QoS specified in SLA according to deadline constraints. So above mentioned two approaches is beneficial to cloud providers in terms of power consumption. Cloud computing also uses the concept of virtualization which enables combining several physical systems into virtual machines on one single powerful system. Therefore unplugging the original hardware & reducing power in cooling consumption.

5. STORAGE (DATA CENTER)

Green data centers doesn't just save energy, they also reduce need for expensive infrastructure upgrades to deal with increased power and cooling demands. Data centers holds data on different pool of servers. Power management is achieved indirectly by detecting idle servers which can be turned off automatically. Distributed Resource Scheduler mechanism will dynamically allocated workloads between physical servers that are created as single resource pool. The system makes adjustments dynamically. In this way, workloads might be consolidated during off-hours and then reallocated across more physical machines as activity increases.

Another option is to categorize data that leads to reduction in power consumption. It includes different kind of data which is accessed by user applications and that data can be categorized as follow

- **Frequently Accessed**

Data that are accessed frequently can be stored on high speed, expensive devices that consume more power

- **Less Frequently Accessed**

Data that is accessed less frequently can be stored on lower speed, less expensive devices that consume less power.

- **Rarely Accessed**

Rarely accessed application and data can be migrated to archival storage devices that result in lowest cost and require lowest power consumption.

5.1 SOLAR POWERED DATA CENTERS

Solar energy is another one of the natural unlimited resource. Power cost can also be lowered by using solar power as electric power in Data Center. IBM'S India software Lab [1] in Bangalore has developed a system to run data centers on solar power and is making it commercially available. A solar power array has been installed spread over more than 6,000 sq.ft. of the lab's rooftop. It is capable of providing 50 kw. of electricity for upto 330 days a year for an average five hours a day. The main advantage of solar power is that it is DC (Direct Current) unlike grid power that is AC (Alternating Current).

Processor run on DC, so when we use grid power. We need to convert AC to DC. In the process of that conversion it loses about 13% of power. On the other hand, thus we can save power close to 10%. In addition; files that are to be accessed regularly are stored in a different set of capacity optimized hard disk. These hard disks enter a low-power mode when not in use and consume negligible energy. To reflect these gains in energy efficiency, our analysis attributes storage power only to those files that are being regularly accessed .While infrequently used data files are stored on a disk.

During periods of low demands, some of the servers enter a sleep mode which reduces a energy consumption. To reflect the efficiency gains from sleeping, virtualization and consolidation, in our analysis, the computation servers and file hosting servers are fully utilized.

6 .CONCLUSION

In this paper, major focus on crisis of power consumption by developing efficient algorithms using sharp logic and efficient data structures. It helps in to finish tasks quickly which leads to free up the computer resources and turned it to off mode. As a result amount of power consumption becomes less. Power management can also be managed through hardware interfaces like ACPI & DVFS. DVFS reduces operating voltage to the minimum level as needed and ACPI controls amount of power required to different resources of computer through interacting with operating system.

The latest emerging paradigm of cloud computing facilitates to both cloud providers and users. Cloud environment allows provision of virtual machines

and services through different techniques without affecting performance. Cloud computing also uses the concept of virtualization which enables combining several physical systems into virtual machines on one single powerful system. Therefore unplugging the original hardware & reducing power in cooling consumption. Green data centers doesn't just save energy, they also reduce need for expensive infrastructure upgrades to deal with increased power and cooling demands. Power management is achieved indirectly by detecting idle data servers which can be turned off automatically. Solar energy is another one of the natural unlimited resource. Power cost can also be lowered by using solar power as electric power in Data Center.

7. REFERENCES

[1] http://articles.economictimes.indiatimes.com/2011-11-01/news/30346054_1_solar-power-data-centre-grid-power

[2] http://www.acpi.info/

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

[4] T.D.Burd and R.W. Brodersen. Enegy efficient cmos microprocessor design. In proc. Of Annual Hawaii Intl.Conf. on system sciences, pages 288-297, January 1995.

[5]M.Cardosa, M.R.Korupolu and A.Singh. Shares and utilities based power consolidation in virtualized server environments. In Proc.of IFIP/IEEE Intl.Symp. on Integrated Network Management .USA, June 2009.

[6] K.H.Kim, A.Beloglazov and Rajkumar Buyya. Power aware provisioning of cloud resources for real time services .In proc.of MGC'09 ,Urbanan Champaign,Illinois,US.

[7] VirtualLogicx Real-Time Virtualization and VLX

VirtualLogicx, http://www.oswara.com

[8]Verma A, Ahuja P, Neogi A. pMapper :Power and Migration cost aware application placement in virtualized systems.In Pro. Of 9th ACM /IFIP/USENIX International conference on middleware ,Leuven, Belgium, December 2008.

[9]Verma A, Ahuja P, Neogi A. Power aware dynamic placement of HPC applications.In Proc.of the 22nd ACM International Conference on supercomputing (ICS 2008). Agean sea, Greece, June 2008.

[10]Amazon elastic compute cloud (Amazon EC2).http://aws.amazon.com/ec2

[11] http://en.wikipedia.org/wiki/Euclidean_algorithm

[12] www.eecs.berkeley.edu/~vazirani/s99cs170/n

IT 043

A Look into Density Based Clustering Algorithms

Ms. Roopa Praveen

Associate Professor

ASM's IBMR, Chinchwad, Pune

Abstract

This paper provides an insight into Density based Clustering Algorithms. The paper focus not only on renowned density based algorithms such as DBSCAN ,DBCLASD, GDBSCAN, OPTICS, DENCLUE but also algorithms like PDBSCAN, DENCLUE2.0,UDBSCAN, which have better efficiency and cluster clarity even in varying conditions and densities. Here not only advantages and disadvantages of the most common algorithms are discussed but also complexity and run time is given.

Keywords:

Density based Clustering, Clustering algorithm, Clustering in the Presence of Noise.

INTRODUCTION

Clustering is the technique which is used in almost every place in our real word and different clustering techniques are used to implement clustering .Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics .To give operational definition is always tough rather than to give functional definition. Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such region containing a relatively low density of points. Clustering's main objective is to distribute objects, events etc in groups such that degree of association should be strong enough among members of same cluster and should be weak among members of different clusters.

Generally clustering is classified into two categories i.e. non-exclusive (overlapping) and exclusive (non-overlapping). Exclusive clustering is further divided into two categories i.e. extrinsic (supervised) and intrinsic (unsupervised). Now intrinsic clustering is further divided into hierarchical

and partitional methods. Our main concern, density based algorithms belong to partitional methods. We will present a brief introduction of both methods.

Hierarchical clustering, as its name suggests is a sequence of partitions in which each partition is nestled into the next partition in the sequence i.e. Maintaining hierarchy. Hierarchical clustering is depicted by binary tree or dendrogram. Whole data set is represented by root node and rest of the leaf node is represented as data object. At any level cutting a dendrogram defines clustering and identifies new clusters. Partitional clustering is of non hierarchical type.

DENSITY BASED CLUSTERING

In Density based clustering there is partition of two regions i.e. low density region to high density region .A cluster is defined as a connected dense component that grows in any direction where a density leads. This is the reason that density based algorithms are capable of discovering clusters of arbitrary shapes and provides natural protection to outliers. Basically, density based clustering is divided into two categories i.e. density based connectivity and density function. While discussing about density based connectivity, density and connectivity are two main concept comes under this and both measured in terms of local distribution of nearest neighbours. Density based connectivity includes DBSCAN, GDBSCAN, OPTICS and DBCLASD algorithms and density function includes DENCLUE algorithm.

The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering; it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius.

1. DBSCAN

In DBSCAN (Density-Based Spatial Clustering of Applications with Noise) it relies on a density based notion of cluster and discovers the clusters and noise in a database and based on fact that a cluster of arbitrary shape with noise.

Now two basic concepts or definitions are discussed here which will be used in many of density based algorithms and in DBSCAN as well.

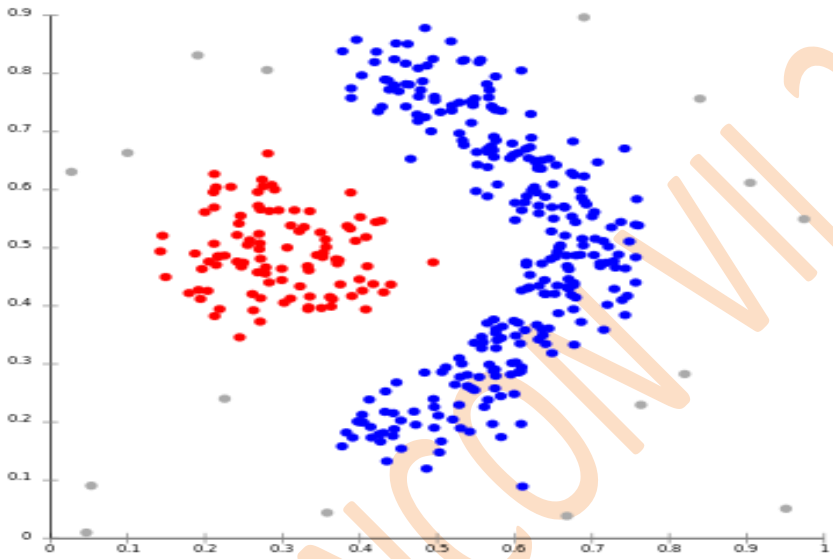
Density reachable-A point p is density reachable from a point q with respect to Eps, Minpts if there is a chain of points p1…..pn such that pi+1 is directly density reachable from pi .

Density connected- A point p is density connected to a point q with respect to Eps, Minpts if there is a point m such that both p and q are density reachable from pi

Selection of cluster is done but validation of cluster is still remaining. Following lemmas are used for this purpose i.e. clusters satisfying following lemmas will be considered as validates cluster and lemmas are as follows:

Lemma 1: Let p be a point in D and INEps(p)I> MinPts. Then the set M = {m I m Є D and m is density-reachable from p wrt. Eps and MinPts} is a cluster wrt. Eps and MinPts.

Density-based clustering with DBSCAN



Lemma 2: Let C be a cluster wrt. Eps and MinPts and let p be any point in C with INEps(p)I> MinPts Then C equals to the set M = {mI m is density-reachable from p wrt. Eps and MinPts}.

The runtime of the algorithm is of the order O (n log n) if region queries are efficiently supported by spatial index structures, i.e. at least in moderately dimensional spaces.

(1) Advantages

a) DBSCAN does not require you to know the number of clusters in the data a priori, as opposed to k-means.

b) DBSCAN can find arbitrarily shaped clusters. It can even find clusters completely surrounded by (but not connected to) a different cluster. Due to the MinPts parameter, the so-called single-link effect (different clusters being connected by a thin line of points) is reduced.

c) DBSCAN has a notion of noise.

d) DBSCAN requires just two parameters and is mostly insensitive to the ordering of the points in the database.

(2) Disadvantages

a) Need to specify global parameters Eps, MinPts in advance from user, which is very difficult.
b) DBSCAN does not respond well to data sets with varying densities (called hierarchical data sets) .

2. UDBSCAN

UDBSCAN (Density based clustering for uncertain objects) .Existing traditional clustering algorithm were designed to handle static objects. The UDBSCAN which is an extension of DBSCAN algorithm uses sample based on the object's uncertain model.
In UDBSCAN extends the existing DBSCAN algorithm to make use of their derived vector deviation function which defines deviation in each direction from the expected representative. Vector deviation, VD= (vd1,……,vd2m), is a set of 2m vectors which initiate from the expected representative. In this a new metric is also defined to measure the quality of cluster of density based clustering.

(1) Advantage

a) Overcome the drawback of DBSCAN and used for uncertain objects.

(2) Disadvantage

a) Problem with this algorithm is that it finds very difficult to extend to high dimensional spaces.

3. GDBSCAN

GDBSCAN (Generalized Density-Based Spatial Clustering of Applications with Noise).It is a generalized version of DBSCAN. It can cluster point objects as well as polygon objects using spatial and non-spatial attributes. GDSCAN generalized DBSCAN in two ways; first if symmetric and reflexive are the two properties of neighbourhood then we can use any notion of the neighbourhood of an object. The two properties that are symmetric and reflexive are termed as binary predicate, second by calculating the non spatial attributes by defining cardinality of the neighbourhood. GDBSCAN has five important applications. In the first application we cluster a spectral space (5D points) created from satellite images in different spectral channels which is a common task in remote sensing image analysis. The second application comes from molecular biology. The points on a protein surface (3D points) are clustered to extract regions with special properties. To find such regions is a subtask for the problem of protein-protein docking. The third application uses astronomical image data (2D points) showing the intensity on the sky at

different radio wavelengths. The task of clustering is to detect celestial sources from these images. The last application is the detection of spatial trends in a geographic information system. GDBSCAN is used to cluster 2D polygons creating so-called influence regions which are used as input for trend detection.

Spatial index structures such as R-trees may be used with GDBSCAN to improve upon its memory and runtime requirements and when not using such a structure the overall complexity is O (n log n).

4. P-DBSCAN

P-DBSCAN, a new density-based clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. Here two new concepts are introduced:
(1) Density threshold, which is defined according to the number of people in the neighbourhood, and
(2) Adaptive density, which is used for fast convergence towards high density regions.

5. OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure).The basic idea of OPTICS and DBSCAN algorithm is same but there is one drawback of DBSCAN i.e. while densities are varying it is difficult to detect meaningful clusters which was overcome in OPTICS algorithm. It not only stores the core distance but also a suitable reachability distance for each object and creates a proper ordering of database. It requires two parameters ε and MinP. E describes radius and MinP tells minimum no of points that are required to form a cluster. Point p is a core point if at least MinP are found within its ε neighbourhood Nε(p).

Core_distε,MinP(p)= Distance to the MinPth point,otherwise
undefined if (|Nε ' (o)|>=MinP)
Reachability_distε ,MinP(p,o) = Max(core_ distε ,MinP(o), dist(o,p) ,othrw

(1) Advantage

a) Clustering ordering can be used to extract basic clustering information.

(2) Disadvantage

b) Wide range of parameter setting is required.
Complexity is O(kN2) where k is no of dimensions Run time is O(n log n).

6. DBCLASD

DBCLASD (Distribution Based Clustering of Large spatial Databases).The main idea behind DBCLASD is the assumption that points within a given cluster are uniformly distributed i.e points of a cluster are distributed in such a manner that it should be like a homogeneous i.e Poisson point process which is controlled to certain part of the data space just like a rain fall .So according to DBCLASD definition of cluster which is based on the distribution of nearest neighbour distance set (NNDistSet (K)) is mentioned as follows:

Definition Let R be set of points .A cluster K is a nonempty subset of R with following properties:

- NNDistSet(K)(Nearest neighbour distance set of a set of points)has the expected distribution with a required satisfying level.
- Maximality condition: K is maximal i.e. extension done by neighbouring points of K does not fulfill condition (1).
- Connectivity condition: There is connectivity or path for each pair of points (l,m) of the cluster i.e K is connected.

Basically two steps are followed in DBCLASD which are as follows:

- Generating Candidates

First step is to generate candidates i.e. for each new member p of cluster K ,we retrieve new candidate using a region query (i.e. circle query) by selected radius which is choosen such that for no point of cluster a larger distance to the nearest neighbour is expected. So a necessary condition for m(radius) is $N*P(NNDistK(p)>m)<1$.

(1) Advantages

a) Better run time than CLARANS
b) DBCLASD requires no user input

(2) Disadvantages

a) It is slower than DBSCAN
The run time of DBCLASD is roughly three times the run time of DBSCAN

7. DENCLUE

DENCLUE (Density based Cluslering).In this algorithm concept of influence and density function is used .Here according to authors influence of each data point can be modelled formally using a mathematical function and that is called an influence function. Influence function describe the impact of data point within its neighbourhood. Now the next concept comes of density function which is sum of influences of all data points.

According to DENCLUE two types of clusters are defined i.e. centre defined and multi centre defined clusters .In centre defined cluster a density attractor x* ( ) is the subset of the database which is density attracted by x* and in

multicenter defined cluster it consist of a set of center defined clusters which are linked by a path with significance ξ and ξ is noise threshold.
The influence function of a data object y Є Fd is a function fY : Fd → R+0 which is defined in terms of a basic influence function Fb fB (x) =- fB (x, y).
fyB (x) =- fB (x, y).
The density function is defined as the sum of the influence functions of all data points.
DENCLUE also generalizes other clustering methods such as density based clustering; partition based clustering, hierarchical clustering. In density based clustering DBSCAN is the example and square wave influence function is used and multicenter defined clusters are here which uses two parameter σ = Eps, ξ = MinPts. In partition based clustering example of k-means clustering is taken where Gaussian influence function is discussed. Here in center defined clusters ξ=0 is taken and σ is determined. In hierarchical clustering center defined clusters hierarchy is formed for different value of σ.

Faster than DBSCAN by a factor of up to 45.
(1) Advantages

a) It has a firm mathematical basis.
b) It has good clustering properties in data sets with large amounts of noise.
c) It allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets.
d) It is significantly faster than existing algorithms

(2) Disadvantages

a) Data points are assigned to clusters by hill climbing, the used hill climbing may make unnecessary small steps in the beginning and never converges exactly to the maximum, it just comes close.
b) Needs a large no of input parameters

8. DENCLUE

DENCLUE2.0 (Fast Clustering based on Kernel
Density Estimation) [13].In both DENCLUE1.0 and DENCLUE2.0 hill climbing procedure is used. Here the main aim of this hill climbing procedure is to maximize the density i.e. ^p(x). Now in gradient based hill climbing first derivative of ˆp(x)is set to be zero and solve for x. Resulting equation is

$$X = \frac{\sum_{t=1}^{N} K(x-x_t/h)x_t}{}$$

$$\sum_{t=1}^{N} K(x - xt / h)$$

Where xt ЄX Rd ,d Є N, t = 1….N and K = Gaussian Kernel

Since x influences the right hand side only through the kernel, the idea is to compute the kernel for some fixed x and update the vector on the Left hand side according to above formula. This gives a new iterative procedure with the update formula

$$X = \frac{\sum_{t=1}^{N} K((x(l) - x(t))/h)xt}{\sum_{t=1}^{N} K(x(l) - x(t))/h}$$

Where h = quantity

CONCLUSION:

A. **Summary**

This paper presents an up-to-date on density based clustering algorithms. It tries to focus not only the renowned algorithms such as DBSCAN,GDBSCAN, DBCLASD, OPTICS ,DENCLUE but also on algorithmssuch as UDBSCAN, PDBSCAN, DENCLUE2.0 that having improvements in efficiency and runtime than existing algorithms. In addition advantages and disadvantages of the most common algorithms with run time and complexities are discussed.

B. **Future Work**

The entire density based clustering algorithms are efficient but the question arises what are the scenarios in which these algorithms performs better among themselves .More generally we can say that which algorithm is more computationally efficient when compared with others.

REFERENCES:

1. Pavel Berkhin, Survey of Clustering Data Mining Techniques , Wyswilson,2002
2. Ester M., Kriegel H.-P., Xu X.: ―Knowledge Discovery in Large Spatial Databases: Focusing
3. Techniques for efficient Class Identification‖, Proc. 4th Int. Symp. on large Spatial Databases,

4. Portland, ME, 1995, in: Lecture Notes In Computer Science, Vol. 951, Springer, 1995, pp. 67-82
5. SANDER, J., ESTER, M., KRIEGEL, H.-P., and XU, X. 1998. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. In Data Mining and Knowledge Discovery, 2, 2, 169-194
6. P.N. Tan, M. Steinbach,V. Kumar, Introduction to Data Mining,Pearson , Addison Wesley,2006
7. S.Kisilevich, F. Mansmann, D. Keim, —P-DBSCAN: a density based Clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos in Proc of COM,Geo _ 10 of the 1st international conference and exhibition on computing for geospatial Research & Application, doi>10.1145/1823854.1823897
8. A. K. Jain, M. N. Murty and P. J. Flynn, Data clustering: a review, ACM, 31 (1999), pp. 264–323

IT 044

E Business and Service Sector

MR. UJJVAL S. MORE

Assistant Professor (Computer Management),

ASM's Institute of Business Management & Research (IBMR),

Chinchwad, Pune Maharashtra INDIA 411019

Email: ujjval.more@asmedu.org

Mobile : 9423497871, 7276545621

Abstract:

In the few years since 1995, electronic commerce has grown in the United States from a standing start to $60 billion retail business and a $700 billion business-to-business, bringing about enormous change in the business firm around globe, in Europe, Asia and Latin America, are being similarly affected. In the next five years, e-commerce in all of its forms is projected to continue growing at double rates becoming the fastest growing form of commerce in the world. The rapid movement towards on e-commerce economy and society is being led by established business firms such as Wal-Mart, General Electric and eBay.

The business services deliveries services aimed on the business life. This group consist of professional services on the area of law, accountancy, accounting and administration, recruitment, marketing, computer sector and technical services such as architect and engineers, real estate market and rental companies that do not deliver touchable material but deliver services to other companies.

Keywords: e-commerce, e-business, B2C, B2B, C2C, P2P, M-commerce, e-commerce – , e – commerce –II, business model, business plan, portal, e-tailor, content provider, transaction broker, FIRE, financial portals, online travel.

Introduction:

E-Commerce means the use of internet and the web to transact the business. More formally we focus on digitally enabled commercial transactions between and among organizations individuals. Digitally enables transaction include all transactions mediated by digital technology. This means transaction that occurs over Internet and on the web. Commercial transaction involve the exchange of value (money) across organizational or individual boundaries in return for the product.

E-Business means the digital enablement of transactions and process within a firm, involving information systems under the control of the firm. E-Business does not include commercial transactions involving and exchange of value across organizational boundaries.

Seven Unique features of E-Commerce Technology:

- Ubiquity – Internet/ Web technology is available everywhere: at work, at home, and elsewhere via mobile devices, anytime.
- Global Reach- The technology reaches across national boundaries, around the earth.
- Universal Standards - There is one set of technology standards, namely Internet standards.
- Richness-Video, audio, and text messages are possible.
- Interactivity - The technology works through interaction with the user.
- Information Density- The technology reduces information costs and raises quality.
- Personalization / Customization- The technology allows personalized messages to be delivered to individuals as well as groups.

Types of E-Commerce:

B2C. Business–to-Consumer (B2C), in which online business attempt to reach individual consumers. Seven different B2C business models are portals, online retailers, content providers, transaction brokers, market creators, service providers, and community providers.

Example – Amazon.com is a general merchandiser that sells consumer products to retail consumers.

B2B. (Business-to-Business (B2B) e-commerce, in which businesses focus on selling to other businesses. The ultimate size of B2B e-commerce could be huge. At first, B2B e-commerce primarily involved inter-business exchanges, but a number of other B2B business models have developed, including e-distributors, B2B service providers, matchmakers, and infomediaries that are widening the use of B2B

Example*:* esteel.com is a steel industry exchange that creates and electronic market for steel producers and users.

C2C. (Consumer-to-Consumer (C2C) e-commerce provides a way for consumers to sell to each other, with the help of an online market maker. In *C2C* e-commerce, the consumer prepares the product for market, places the product for auction or sale, and relies on the market maker to provide catalog, search engine, and transaction-clearing capabilities so that products can be easily displayed, discovered, and paid for.

***Example*:** eBay.com creates a marketplace where consumers can auction or sell goods.

P2P. (Peer-to-Peer) This technology enables Internet users to share files and computer resources directly without having to go through a central Web Server. In Peer-to-Peer's purest form, no intermediary is required.

***Example*:** Gnutella is a software application that permits consumes to shape music with one another directly, without the intervention of a market maker as in C2C e-commerce.

M-Commerce (Mobile Commerce) This technology uses the wireless digital devices to enable the transaction on the Web. They utilize wireless networks to connect cell phones.

***Example*:** PDA or mobile can used for stock price, store prices compilation banking, travel reservations and more.

History of E-Commerce can be divided into two periods:

E-*Commerce – I: A* period of explosive growth in e-commerce beginning in 1995 and ending in 2000.

E-Commerce – II: The current era of e-commerce, beginning in 2001.

Table 1 E - Commerce – I and E – Commerce Comparison

Technology-driven Revenue	Earnings and profits emphasis
growth emphasis Venture	Traditional financing
capital financing	Stronger regulation and governance
Ungoverned Entrepreneurial	Large traditional firms Strengthening
Disintermediation	intermediaries
Perfect markets	Imperfect markets, brands, and network effects
Pure online strategies	Mixed" clicks and bricks "strategies
First mover advantages	Strategic follower strength

Success of E- Commerce – I

- A technological success, with the digital infrastructure created during the period solid enough to sustain significant growth in e-commerce during the next decade.
- A mixed business success, with significant revenue growth and customer usage, but low profit margins.

Failure of E- Commerce – I

- Fulfilled economists' visions of the perfect Betrand market and friction-free commerce.

- Fulfilled the visions of entrepreneurs and venture capitalists for first mover advantages, low customer acquisition and retention costs, and low costs of doing business.

Factors of e - commerce II era

- E-commerce technology will continue to propagate through all commercial activity, with overall revenues from e-commerce, the number of products and services sold over the Web, and the amount of Web traffic all rising.
- E-commerce prices will rise to cover the real costs of doing business on the Web.
- E-commerce margins and profits will rise to levels more typical of all retailers.
- Traditional well-endowed and experienced Fortune 500 companies will playa growing and more dominant role.
- The number of successful pure online companies will continue to decline and most successful e-commerce firms will adopt a mixed "clicks and bricks" strategy.
- Regulation of e-commerce and the Web by government will grow both in the United States and worldwide.

The major themes underlying the study of E-commerce:

- **Technology** *:* The understand e-commerce, you need a basic understanding of the information technologies upon which it is built, including the Internet and the World Wide Web, and a host of complimentary technologies - personal computers, local area networks, client/server computing, packet-switched communications, protocols such as TCP/IP, Web servers, HTML, and relational databases, among others.
- **Business***:* While technology-provides the infrastructure, it is the business applications - the potential for extraordinary returns on investment - that create the interest and excitement in e-commerce. New technologies present businesses and entrepreneurs with new ways of organizing production and transacting business. Therefore, you also need to understand some key business concepts such as electronic markets, information goods, business models, firm and industry value chains, industry structure, and consumer behavior in electronic markets

- **Society**: Understanding the pressures that global e-commerce place on contemporary society is critical to being successful in the e-commerce marketplace. The primary societal issues are intellectual property, individual privacy, and public policy.

The major academic discipline contributing to e-commerce:

- Computer scientists are interested in e-commerce as an application of Internet technology.
- Management scientists are primarily interested in building mathematical models of business processes and optimizing them to learn how businesses can exploit the Internet to improve their business operations.
- Information systems professionals are interested in e-commerce because of its implications for firm and industry value chains, industry structure, and corporate strategy.

- Economists have focused on consumer behavior at Web sites, and on the features of digital electronic markets.
- Sociologists have focused on studies of Internet usage, the role of social inequality in skewing Internet benefits, and the use of the Web as a personal and group communications tool.
- Finance and accounting scholars have focused on e-commerce firm valuation and accounting practices,
- Management scholars have focused on entrepreneurial behavior and the challenges faced by young firms who are required to develop organizational structures in short time spans,
- Marketing scholars have focused on consumer response to online marketing and advertising campaigns, and the ability of firms to brand, segment markets, target audiences, and position products to achieve higher returns on investment.

E – Commerce Business Model:

Business model is a set of planned activities designed to result in a profit in market place.
Business plan is a document that describes a firm's business model.
Revenue model describes how the firm will earn revenue, produce profits and produce a superior return on invested capital.
Advertising revenue model a company provides a forum for advertisements and receives fees from advertiser.

Subscription revenue model a company offers its users content or service and charges a subscription fee for access to some or all of offerings.

Transaction fee revenue model a company receives a fee for enabling or executing a transaction.

Sales revenue model a company derives revenue by selling goods, information or services.

Affiliate revenue model a company steers business to an affiliate and receives a referral fee or percentage of the revenue from any resulting sales.

Table 2 Five Primary revenue models

Revenue Model	***Example***	***Revenue Source***
Advertising	Yahoo.com	Fees from advertisers in exchange for advertisements
Subscription	WSJ.com, Consumerreports.org, Sportsline.com	Fees from subscribers in exchange for access to content or services
Transaction Fee	eBay.com, E- Trade.com	Fees (commissions) for enabling or executing a transaction
Sales	Amazon.com, DoubleClick.net, Salesforce.com	Sales of goods, information, or services
Affiliate	MyPoints.com	Fees for business referrals

Table 3 B2C Business Models

Portal	Horizontal / General	Yahoo. com, AOL.com,	Offers an integrated package of services and content such	Advertising, subscription

		MSN.com, Excite@home.com	as search, news, e-mail, chat, music downloads, video streaming, and calendars. Seeks to be a user's home base.	fees, transaction fees
	Vertical / Specialized	iBoats.com	Offers services and products to specialized marketplace.	Advertising, subscription fees, transaction fees
E-tailer	Virtual Merchant	Amazon.com	Online version of retail store, where customers can shop at any hour of the day or night without leaving home or office.	Sales of goods
	Clicks and Mortar	Walmart.com	Online distribution channel for company that also has physical stores.	Sales of goods
	Catalog Merchant	LandsEnd.com	Online version of direct mail catalog.	Sales of goods
	Online Mall	Fashionmall.com	Online version of mall.	Sales of goods, transaction fees
	Manufacturer- direct	Dell.com	Online sales made directly by manufacturer.	Sales of goods
Content Provider		WSJ.com, Sportsline.com, CNN.com	Information and entertainment providers such as newspapers, sports sites, and other online sources that offer customers up- to-date news and special interest, how-to guidance, and tips and/or information sales.	Advertising, subscription fees, affiliate referral, fees
Transaction Broker		E- Trade.com, Expedia.com, Monster.com	Processors of online sales transactions, such as stock brokers and travel agents, that increase customers' productivity by helping them get things done faster and	Transaction fees

			more cheaply.	
Market Creator	Auctions and other forms of dynamic pricing	eBay.com, Priceline.com	Web-based businesses that use Internet technology to create markets that bring buyers and sellers together.	Transaction fees
Service Provider		xDrive.com, whatsitworthtoyou.com, myCFO.com	Companies that make money by selling users a service, rather than a product.	Sales of services
Community Provider		About.com, iVillage.com, BlackPlanet.com	Sites where individuals with particular interests, hobbies, and common experiences can come together and compare notes.	Advertising, subscription, affiliate referral fees

Table 4 B2B Businesses Model:

BUSINESS MODEL	**VARIATIONS**	**EXAMPLES**	**Description**	**Revenue Model**
Marketplace / Exchange (B2B Hub)	Vertical	DirectAg.com, e-Steel.com	Helps bring buyers and sellers together to reduce procurement costs for a specific industry	Transaction Fees
	Horizontal	TradeOut.com	Same as vertical except focused on specific types of products and services.	Transaction Fees
E-Distributor		Grainger.com	Connecting businesses directly with other businesses, reducing sales cycles and mark-up.	Sales of goods

B2B Service Provider	Traditional	Employeematters.com	Supports companies through online business services.	Sales of services
	Application Service Provider (ASP)	Salesforce.com, Corio.com	Rents Internet-based software applications to businesses.	Rental fees
Matchmaker		iShip.com	Helps business find what they want and need on the web.	Referral fee
Infomediary	Audience Broker	DoubleClick.net	Gathers information about consumers and uses it to help advertisers find the most appropriate audience	Sales of Information,
	Lead generator	AutoByTel.com	Gather information about	Referral fee

•

Table 5 Businesses Models in Emerging E-Commerce Areas

Type	***Model***	***Example***	***Description***	***Revenue Model***
Consumer-to-consumer	Market Creator	eBay.com, Half.com	Helps consumers connect with other consumers who have items to sell.	Transaction fees
Peer-to-peer	Content Provider	Napster.com, My.MP3.com	Technology enabling consumers to share files and services via the Web.	Subscription fees, advertising, transaction fees

M-commerce	Various	Amazon.com	Extending business applications using wireless technology.	Sales of goods

•
•

No discussion of e-commerce business models would be complete without mention of a group of companies whose business model is focused on providing the infrastructure necessary for e-commerce companies to exist, grow, and proper

Table 6 E – Commerce Enablers

Infrastructure	***Players***
Hardware: Web Servers	IBM, Sun, Compaq, Dell
Software: Operating Systems and Server Software	Microsoft, Sun, Apache Software Foundation
Networking: Routers	Cisco
Security: Encryption Software	CheckPoint, VeriSign
E-commerce Software Systems (B2C, B2B)	IBM, Microsoft, iPlanet, CommerceNet, Ariba
Streaming Media Solutions	Real Networks, Microsoft
Customer Relationship Management Software	PeopleSoft
Payment Systems	PayPal, CyberCash
Performance Enhancement	Akamai, Cache Flow, Inktomi, Cidera, Digital Island
Databases	Oracle, Sybase
Hosting Services	Exodus, Equinex, Global Crossing

Major feature of Online Service Sector:

Service sector is the largest and most rapidly expanding part of the economy of advanced industrial nations. Service industries are companies that provide services e., perform tasks for) consumers, businesses, governments, and other organizations. The major service industry groups are FIRE, business services, and health services. Within these service industry groups, companies can be further categorized into those that involve transaction brokering and those that involve providing "hands-on" service. With some exceptions, the service sector is by and large a knowledge and information-intense industry. For this reason, many services are uniquely suited to e-commerce and the strengths of the Internet. The rapid expansion of e-commerce services in the areas of finance, including insurance and real estate, travel, and job placement, can be explained by the ability pf these firms to:

- collect, store and disseminate high value information ;
- provide reliable, fast communication; and
- personalize and customize service or components of service.

E-commerce offers extraordinary opportunities to improve transaction efficiencies and thus productivity in a sector where productivity has so far not been markedly affected by the explosion in information technology.

Trends taking place in the online financial services industry:

The online financial services sector is a good example of an e-commerce success story, but the success is somewhat different than what had been predicted in the E· commerce I era, Pure online financial services firms are in general not yet profitable. In E-commerce II, once again, it is the multichannel established financial firms that are growing the most rapidly and that have the best prospects for long· term viability. Other significant trends include the following:

- Management of financial assets online is growing rapidly. By 2005, it is projected that more than half of all households in the United States will be investing online; online banking is also expected to more than double by 2005: online stock trading is expected to grow from 15 million households today to 34 million households in 2005.
- In the insurance and real estate industries, consumers still generally utilize the Internet just for research and use a conventional transaction

broker to complete the purchase.

- Historically, separate institutions have provided the four generic types of services provided by financial institutions. Today, as a result of the Financial Reform Act of 1998, which permitted banks, brokerage firms, and insurance companies to merge, this is no longer true. This has resulted in two important and related global trends in the financial services industry that have direct consequences for online financial services firms: the move toward industry consolidation and the provision of integrated financial services.

The major trends in Online Travel Services:

The major trends include the following:

- The online travel services industry is going through a period of consolidation as stronger offline, established agencies purchase weaker and relatively inexpensive online travel agencies in order to build stronger multichannel travel sites that combine physical presence, television sales outlets, and online sites.
- Suppliers - such as airlines, hotels, and auto rental firms - are attempting to eliminate intermediaries such as GDS and travel agencies, and develop a direct relationship with consumers; an example is Orbitz, the online reservation system formed by seven major airlines to deal directly with corporations and leisure travelers. The major auto rental firms have also all opened direct-to-customer Web sites. At the same time, successful online travel agencies are attempting to turn themselves into merchants by purchasing large blocks of travel inventory and then reselling it to the public, eliminating the global distributors and earning much higher returns.

Why online career services may be the Ideal Business?

Job hunting services have been one of the Internet's most successful online services because they save money for both job hunters and employers. In comparison to online recruiting, traditional recruitment tools have severe limitations.

- Online recruiting provides a more efficient and cost-effective means of linking employers and job hunters and reduces the total time-to-hire.
- Job hunters can easily build, update, and distribute their resumes, conduct job searches, and gather information on employers at their convenience and leisure.
- It is an information-intense business process which the Internet can automate, and thus reduce search time and costs for all parties.

Online recruiting can also serve to establish market prices and terms, thereby identifying both the salary levels for specific jobs and the skill

sets required to achieve those salary levels. This should lead to a rationalization of wages, greater labor mobility, and higher efficiency in recruitment and operations as employers are able to more quickly fill positions.

Current trends in the online career services industry include the following:

- Although online recruitment has focused primarily on general job recruitment, many general sites have begun listing lower- and middle-level management jobs, which offer the largest revenue potential.
- The online recruitment industry is going through a period of rapid consolidation led by TMP Worldwide Inc, which because of its acquisitions, is now the world leader in online executive recruiting, executive moving and relocation, e-resourcing, temporary contracting, and recruitment advertising.

E – Business Pros:

- Lower Start up cost
- International customer Access
- Business conducted 24 X 7
- Quick response to customers
- Lower overhead
- Lower product cost
- Work from home or remotely

E – Business Cons:

- No immediate customer visibility
- You cant see physically customer
- Constant need to maintain sites
- Late nights and long hours
- Lack of Customer trust due to location
- Lack of immediate health care benefits
- Disruption of business and family privacy

Conclusion:

E-business is still quickly gaining popularity among web users young and old. It is particularly appealing to people in remote areas who can now visit their favorite stores online with just the click of the mouse. Shopping online is fast,

convenient and buyers can shop in the comfort of their own home. This is why both buyers and sellers are eager to get into e-commerce and shopping online. Some of the pros can make things more complicated. For instance International customer access means that you have a potentially global customer base, but is also means that you should be aware of the distinction between National and International shipping requirements. But knowing the difference between online & traditional business will help us to maximize the benefits of an e-business and its services.

References:

[1] Kenneth C. Laudon, Carol Guercio Traver '*E – Commerce*' , Pearson Education by Dorling Kindersley (India) Pvt. Ltd, New Delhi, 2008.

[2] Arthur Brian. "Increasing Returns and the New world of Business", *Harvard Business Review* (July – August 1996).

[3] Bailey, Joseph P. *Intermediation of Electronic Markets: Aggregate pricing in Internet Commerce.* Ph. D., Technology, Management and Policy, Massachusetts Institute of Technology (1998a).

[3] Bakos, Yannis. "*The Emerging Landscape for Retail E-Commerce.*" Journal of electronic perspectives (January 2001).

[4] Rayport, Jeffrey and Bernard Jaworski. *e – Commerce.* New York: McGraw – Hill (2000).

[5] Jupiter Media Metrix, (Steven Cutler, Lead Analyst). "*The E * Trade Experience*" (October 2000b).

[6] Laudon Kenneth C., Jane P. Laudon. *Management Information Systems: Managing the Digital Firm,* 7th Edition. Upper Saddle River, NJ: Prentice Hall (2002).

IT 045

SWARM INTELLIGENCE

Preeti Nagnath Whatkar

Student, M.C.A. Dept.

ASM's IBMR College, Pune

Christopher Shashikanth Lagali

Student, M.C.A. Dept.

ASM's IBMR College, Pune

ABSTRACT

"Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial" *by Gerardo Beni and Jing Wang in 1989 in the context of cellular robotic systems.*

To achieve a particular human objective with a technological approach using biological aspects & instincts of living organisms is swarm intelligence.

Since birth human beings are programmed to accomplish certain tasks using a predefined structure or set of rules. As per observations we have seen that certain tasks are out of the range of human beings due to certain constraints like size, temperature, altitude n other atmospheric conditions. These difficult tasks can be accomplished through swarm intelligence.

In other words we are using mechanisms of living organisms that are adapted to certain conditions to accomplish a human defined task.

The study of swarm intelligence is presently confined to laboratory experimentations which have provided insights that can help humans manage complex systems, from truck routing to military robots & aviation industry.

KEYWORDS

Ant colony optimization algorithm, Artificial intelligence, constraints, firefly algorithm, instinct, intelligent water drops algorithm, pheromone, self-organization, social insects, stigmergy, swarm.

INTRODUCTION

Traditional systems were program oriented; systematic who could accomplish tasks that are

- well-defined,
- fairly predictable,
- computable in reasonable time with serial computers.

The demand for intelligent & sensitive systems forced scientists to explore the fields of physics, chemistry & biology to be applied with artificial intelligence to develop sophisticated complex systems that can achieve unpredictable, difficult solutions with fair amount of adaptability.

The need to achieve this objective the following alternatives took birth

- DNA based computing (chemical computation).
- Quantum computing (quantum-physical computation)
- Biologically inspired computing (simulation of biological mechanisms)

Swarm intelligence is part of biologically inspired computing where biological aspects such as instincts, living conditions, social behaviour are implemented along with artificial intelligence. To some extent the intention is to make computer systems that think & behave like some living organisms that functions even under extreme conditions where living organisms cannot exist or tolerate such conditions.

EXAMPLES WHERE SWARM INTELLIGENCE IS IMPLEMENTED

- group foraging of social insects
- co-operative transportation
- division of labour
- nest-building of social insects
- collective sorting and clustering

ANALOGIES IN INFORMATION TECHNOLOGY & SOCIAL INSECTS

- distributed system of interacting autonomous agents
- goals: performance optimization and robustness
- self-organized control and cooperation (decentralized)
- division of labour and distributed task allocation
- indirect interactions

WHAT IS SWARM INTELLIGENCE?

"Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial" *by Gerardo Beni and Jing Wang in 1989 in the context of cellular robotic systems.*

Basic example of Swarm intelligence: ***ANTS***

Ants solve complex tasks by simple local means. Ant productivity is better than the sum of their single activities. Ants are 'grand masters' in search and exploitation. The important mechanisms that ants incorporate are cooperation and division of labour, adaptive task allocation, work stimulation by cultivation & lastly pheromones. The key to an ant colony is that there is no one in command. The entire ant colony functions well without any management at all – at least none that we humans categorize. Instead it relies on the unlimited interactions between the individual ants, each of which is following its own set of rules. Scientists describe such a system as **self - organisation**.

SELF-ORGANISATION

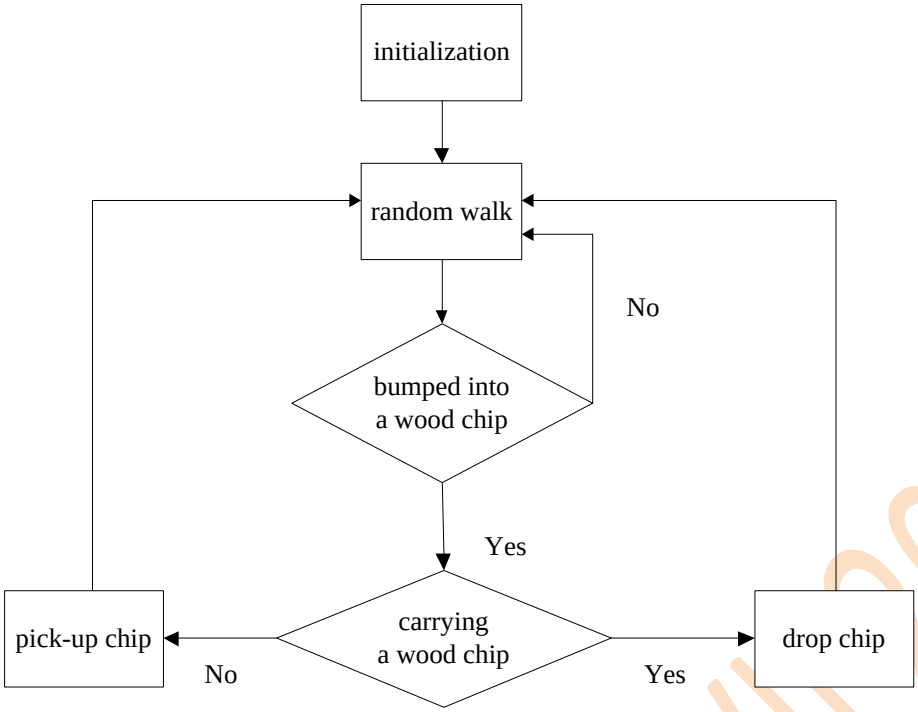
"Self-organization is a set of dynamical mechanisms whereby structures appear at the global level of a system from interactions of its lower-level components."

(Bonabeau et al, in Swarm Intelligence, 1999)

The four bases of self-organisation:

- positive feedback (amplification)
- negative feedback (for counter-balance and stabilization)
- amplification of fluctuations (randomness, errors, random walks)
- multiple interactions

EXAMPLE OF SELF-ORGANISATION IN TERMITE SIMULATION



CHARACTERISTICS OF SELF-ORGANIZED SYSTEMS

- structure emerging from a homogeneous start-up state
- multi-stability - coexistence of many stable states
- state transitions with a dramatical change of the system behaviour

EXAMPLE OF SELF-ORGANIZATION IN HONEY BEE NEST BUILDING

- the queen moves randomly over the combs
- eggs are more likely to be layed in the neighbourhood of brood
- honey and pollen are deposited randomly in empty cells
- four times more honey is brought to the hive than pollen
- removal ratios for honey: 0.95; pollen: 0.6
- removal of honey and pollen is proportional to the number of surrounding cells containing brood

STIGMERGY

Stigmergy: stigma (sting) + ergon (work)

= 'stimulation by work'

The World-Wide Web is the first stigmergic communication medium for humans.

Why does communicating through a weblog work? Stigmergy.

Using a weblog is communicating through stigmergy. Just like an ant, as I blog I leave a trail of information and links to other information I find interesting.

CHARACTERISTICS OF STIGMERGY

- indirect agent interaction modification of the environment
- environmental modification serves as external memory
- work can be continued by any individual
- the same, simple, behavioural rules can create different designs
- according to the environmental state

SOME ALGORITHMS FOR SWARM INTELLIGENCE

1. ANT COLONY OPTIMIZATION

Ant colony optimization (ACO) is a class of optimization algorithms modeled on the actions of an ant colony. ACO methods are useful in problems that need to find paths to goals. Artificial 'ants'—simulation agents—locate optimal solutions by moving through a parameter space representing all possible solutions. Real ants lay down pheromones directing each other to resources while exploring their environment. The simulated 'ants' similarly record their positions and the quality of their solutions, so that in later simulation iterations more ants locate better solutions.

2. FIREFLY ALGORITHM

Firefly algorithm (FA) is another swarm-based algorithm, which was inspired by the flashing behavior of fireflies. Light intensity is associated with attractiveness of a firefly, and such attraction enables the fireflies with the ability to subdivide into small groups and each subgroup swarm around the local modes. Therefore, firefly algorithm is especially suitable for multimodal optimization problems.

3. INTELLIGENT WATER DROPS

Intelligent Water Drops algorithm (IWD) is a swarm-based nature-inspired optimization algorithm, which has been inspired by natural rivers and how they find almost optimal paths to their destination. These near optimal or optimal paths follow from actions and reactions occurring among the water drops and the water drops with their riverbeds. In the IWD algorithm, several artificial water drops cooperate to change their environment in such a way that the optimal path is revealed as the one with the lowest soil on its links. The solutions are incrementally constructed by the IWD algorithm. Consequently, the IWD algorithm is generally a constructive population-based optimization algorithm.

STEPS FOR DESIGNING SI SYSTEMS

1. **identification of analogies**: in swarm biology and IT systems

2. **understanding**: computer modelling of realistic swarm biology
3. **engineering**: model simplification and tuning for IT applications

USE, SCOPE & APPLICATIONS OF SWARM INTELLIGENCE

1. **DEFENCE**

The U.S. military is investigating swarm techniques for controlling unmanned vehicles. Drones are classic examples which use high flying characteristics of the Eagle to carry out reconnaissance and attack missions. These are Unmanned Aircraft Systems (UAV) which are controlled remotely from the nearest Air bases. Chiefly used by the U.S Defence Services in their efforts to pick up intelligence and also carry out Air Strikes. UAVs are also used in a small but growing number of civil applications, such as fire fighting or non-military security work, such as surveillance of pipelines.

2. **MEDICAL FIELD**

A 1992 paper by M. Anthony Lewis and George A. Bekey discusses the possibility of using swarm intelligence to control nanobots within the body for the purpose of killing cancer tumors.

The results of Geoffrey von Maltzahn et al. in their Nature Materials publication reveal that nanoparticles that communicate with each other can deliver more than 40-fold higher doses of chemotherapeutics (anti-cancer drugs) to tumours than nanoparticles that do not communicate can deliver. These results show the potential for nanoparticle communication to amplify drug delivery over that achievable by nanoparticles that work alone, similar to how insect swarms perform better as a group than the individual insects do on their own.

3. **IT SECURITY**

As stated by Dr. Errin Fulp, Associate Professor of Computer Science at Wake Forest University, the digital Swarm Intelligence consists of three components:

Digital ant: Software designed to crawl through computer code, looking for evidence of malware. There could ultimately be 3000 different types of Digital Ants employed.

Sentinel is the autonomic manager of digital ants congregated on an individual computer. It receives information from the ants, determines the state of the local host, and decides if any further action is required. It also reports to the Sergeant.

Sergeant is also an autonomic manager, albeit of multiple Sentinels. Generally, the size of the network would determine how many Sergeants are to be used. Also, Sergeants interface with human supervisors.

Digital ants are simple agents that check for a piece of evidence (malware) and leave pheromone (so other ants can locate the evidence) if malware is found. Sentinels reside on individual computers and interact with ants to discover any threats based on the ants' findings. Sergeants interact with Sentinels and can observe changes over multiple computers.

A file provided by the Sentinel which can be digitally signed to prevent alteration by malware informs other Digital Ants what to focus on. It is a more scalable and robust design. One drawback is speed, as these systems require some time to ramp-up and down. Still it is a worthwhile approach for the massively parallel systems we will face in the future.

4. **AVIATION INDUSTRY**

Airlines have used swarm intelligence to simulate passengers boarding a plane. Southwest Airlines researcher Douglas A. Lawson used an ant-based computer simulation employing only six interaction rules to evaluate boarding times using various boarding methods (Miller, 2010, xii-xviii).

REFERENCES

1. *Ant Colony Optimization* by Marco Dorigo and Thomas Stützle, MIT Press, 2004. ISBN 0-262-04219-3
2. Beni, G.; Wang, J. *Swarm Intelligence in Cellular Robotic Systems*, Proceed. NATO *Advanced Workshop on Robots and Biological Systems*, Tuscany, Italy, June 26–30 (1989)
3. *Bio-inspired computing*, Wikipedia article [Online]
Available: http://en.wikipedia.org/wiki/Bio-inspired_computing
4. Geoffrey von Maltzahn et al. Nanoparticles that communicate in vivo to amplify tumour targeting. *Nature Materials* 10, 545–552 (2011)
5. *International Journal of Bio-Inspired Computation* 1 (1/2): 71–79. [Online]
Available: http://www.inderscience.com/filter.php?aid=22775
6. Lewis, M. Anthony; Bekey, George A. "The Behavioral Self-Organization of Nano robots Using Local Rules". *Proceedings of the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems*
7. Miller, Peter (2010). *The Smart Swarm: How understanding flocks, schools, and colonies can make us better at communicating, decision making, and getting things done*. New York: Avery.
8. Shah-Hosseini, Hamed (2009) [Online]
Available: "The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm"
9. *Swarm intelligence*, Wikipedia article [Online]
Available: http://en.wikipedia.org/wiki/Swarm_intelligence
10. Yang X. S., (2008). *Nature-Inspired Metaheuristic Algorithms*. Frome: Luniver Press. ISBN 1905986106

IT 046

Enabling Non Disruptive Data Migration to Support Continuous Data Availability

Prof. Sachin Patil (Asst. Professor at ASM's IBMR MCA Dept.)

Mob: +91 8087442533. Email: patilsachin4u@gmail.com

Abstract:

The rapidly increasing volumes of data generated by businesses can create significant data management challenges in Medium scale industries. When responding to business demands or compliance and litigation requirements, companies must be able to access and organize volumes of data stored in a variety formats. Maintaining legacy tapes- and the equipment necessary to read them - can be costly and burdensome. In addition, because tape storage has a high failure rate, data recovery is often required in order to restore the data prior to migration. Data migration is necessary when a company upgrades its database or system software, either from one version to another or from one program to an entirely different program. Software can be specifically written, using entire programs or just scripts, to facilitate data migration. Such upgrades or program switches can take place as a result of regular company practices or as a result of directives mandated in the wake of a company takeover. Data Migration is the process of transferring data between storage types, formats, or computer systems. It is often required when organization change computer systems or upgrade to new systems. This solution is usually performed programmatically to achieve an automated migration. As a result, legacy data stored on out-of-date or obsolete formats is evaluated, indexed, de-duplicated and then migrated to newer more cost-effective and reliable storage media. So enabling non disruptive data migration will support continuous data availability and application especially in Medium scale industries which are likely to change.

Keywords: Data Migration, Storage device, Media, Data Cleansing, Data availability.

Introduction:

Data migration means what it sounds like it means -- sort of. It's not data that moves one from place to another, unless you think of places as being virtual. Data migration is actually the translation of data from one format to another format or from one storage device to another storage device. This also necessarily requires someone or something to do the translating. Data doesn't just get up and walk to another format all by itself.

Data migration is necessary when a company upgrades its database or system software, either from one version to another or from one program to an entirely different program. Software can be specifically written, using entire programs or just scripts, to facilitate data migration. Such upgrades or program switches can take place as a result of regular company practices or as a result of directives mandated in the wake of a company takeover. Another use of data migration is to store little-used data on magnetic tape or other backup storage methods. This data may need to be stored for historical purposes or for periodic access. Individual computer users do this all the time when they back up their data to CDs, DVDs, or external hard drives. Companies large and small do this, of course, to protect and archive their data. Migrated data typically is moved offline but remains available via network access, leaving the online environment free to conduct current business.

Need:

The rapidly increasing volumes of data generated by businesses can create significant data management challenges. When responding to business demands or compliance and litigation requirements, companies must be able to access and organize volumes of data stored in a variety formats.

Maintaining legacy tapes—and the equipment necessary to read them—can be costly and burdensome. In addition, because tape storage has a high failure rate, data recovery is often required in order to restore the data prior to migration.

Tape recovery and data migration are one of the most complex activities to manage, predict and perform. Tape media formats need to be read in their native states and imaged to a device which permits low level edits in order to gain access to the information.

Media with physical damage and misaligned write errors all have significant complexity associated with gaining access to the information. The particular agent used and the type of data stored further ads complexity to the scope of the work. A well designed migration methodology ensures that your data is properly evaluated, reviewed and restored before it is migrated to new, more accessible and cost-effective media. And, the process makes certain that your valuable data is safeguarded from beginning to end.

Achieve Data migration:

To achieve an effective data migration procedure, data on the old system is mapped to the new system providing a design for data extraction and data loading. The design relates old data formats to the new system's formats and requirements. Programmatic data migration may involve many phases but it minimally includes data extraction where data is read from the old system and data loading where data is written to the new system.

If a decision has been made to provide a set input file specification for loading data onto the target system, this allows a pre-load 'data validation' step to be put in place, interrupting the standard E(T)L process. Such a data validation process can be designed to interrogate the data to be transferred, to ensure that it meets the predefined criteria of the target environment, and the input file specification. An alternative strategy is to have on-the-fly data validation occurring at the point of loading, which can be designed to report on load rejection errors as the load progresses. However, in the event that the

extracted and transformed data elements are highly 'integrated' with one another, and the presence of all extracted data in the target system is essential to system functionality, this strategy can have detrimental, and not easily quantifiable effects.

After loading into the new system, results are subjected to data verification to determine whether data was accurately translated, is complete, and supports processes in the new system. During verification, there may be a need for a parallel run of both systems to identify areas of disparity and forestall erroneous data loss. Automated and manual data cleaning is commonly performed in migration to improve data quality, eliminate redundant or obsolete information, and match the requirements of the new system.

Data migration phases (design, extraction, cleansing, load, verification) for applications of moderate to high complexity are commonly repeated several times before the new system is deployed.

Some key terms in understanding data migration are:

- ***Legacy data*** is the recorded information that exists in your current storage system, and can include database records, spreadsheets, text files, scanned images and paper documents. All these data formats can be migrated to a new system.
- ***Data migration*** is the process of importing legacy data to a new system. This can involve entering the data manually, moving disk files from one folder (or computer) to another, database insert queries, developing custom software, or other methods. The specific method used for any particular system depends entirely on the systems involved and the nature and state of the data being migrated.
- ***Data cleansing*** is the process of preparing legacy data for migration to a new system. Because the architecture and storage method of new or updated systems are usually quite different, legacy data often does not meet the criteria set by the new system, and must be modified prior to migration. For example, the legacy system may have allowed data to be entered in a way that

is incompatible with the new system. Architecture differences, design flaws in the legacy system, or other factors can also render the data unfit for migration in its present state. The data cleansing process manipulates, or cleans, the legacy data so it conforms to the new system's requirements.

Best practices for data migration (By IBM Global Technology Services)

Data migration is the process of making an exact copy of an organization's current data from one device to another device—preferably without disrupting or disabling active applications—and then redirecting all input/output (I/O) activity to the new device. There are a variety of circumstances that might cause an organization to undertake a data migration, including:

- Server or storage technology replacement or upgrade
- Server or storage consolidation
- Relocation of the data center
- Server or storage equipment maintenance, including workload balancing or other performance-related maintenance.

The above scenarios are fairly routine parts of IT operations in organizations of virtually any size. They are so routine, in fact, that more than 60 percent of respondents to a recent survey1 indicated that they migrate data quarterly or more often—with 19 percent migrating weekly. However, even routine processes can cause problems for IT administrators and managers. More than 75 percent of respondents to the same survey said they had experienced problems during data migration. These problems included, but were not limited to:

- Extended or unexpected downtime
- Data corruption, missing data or data loss
- Application performance issues
- Technical compatibility issues.

How can organizations minimize the business impacts of data migration—downtime, data loss and increased cost? The best way is to employ a consistent, reliable and repeatable methodology for migrations that incorporates planning, technology implementation and validation. Following image shows data migration methodology for best practices.

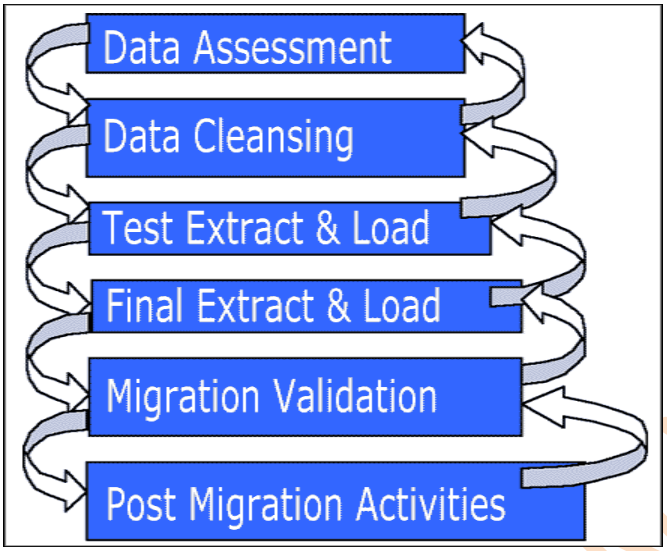


Fig 1

Data migration methodology

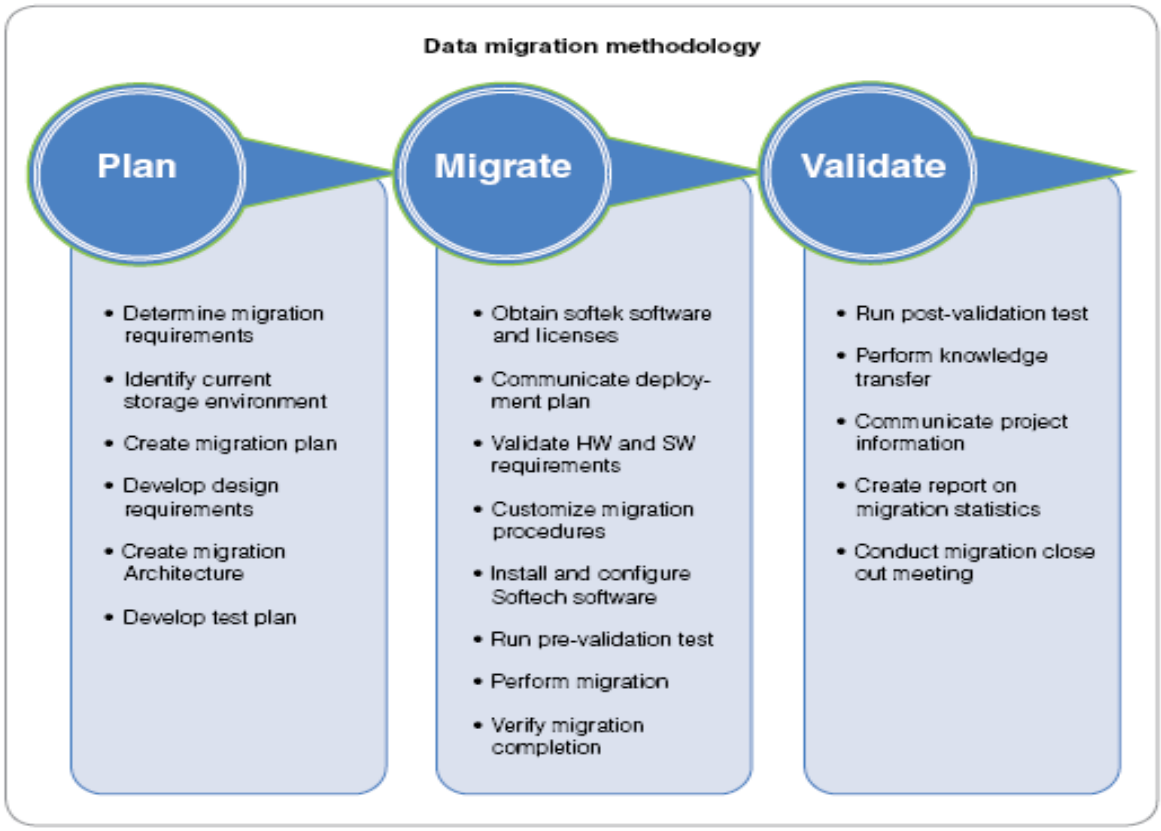


Figure 1: Migration methodology

Setting-up the Data Migration Process:

1. Choose a Data Modeling Tool with Reverse Engineering Capability,(such as ERWin).
2. Define and create the Data Dictionary.
3. Identify all the required Data Sources, and an 'owner' for each Source.

- Data Feeds
- Legacy Systems
- Operational Data Stores

4. If your Sources include Access Databases, then check out MDB Tools, from Brian Burns.
5. Define the Data Items required, in consultation with the Users.
6. Create the Data Models for the Source Data.
7. Define Data Validation Checks (bottom-up) and Clean-Up Business Rules for Source Data.

8. Carry out an Audit of the Data Quality in major Databases, (bottom-up andtop-down). Evaluate the benefits of a Data Cleansing Product, such as Seamless from the C and C Group.
9. Define the Staging Area, with MIRror Tables to store Extract Files.
10. Create the Business Data Model for the Consolidated Database
11. If the final target is an ERP, such as SAP, then create the Data Model for the Target ERP Database.
For SAP, J.D.Edwards, Peoplesoft and Siebel, a Modelling Tool called Saphir, from Silwood Technology can be very useful here.
12. Define the Data Mapping between Source and Target Data Items.
13. Optionally, create a CRUD Matrix to identify the interactions between Data and Functions.
14. Define Acceptance Tests for data in the Integrated Database.

During the migration phase, the IT organization will need to communicate its plans; obtain, install and configure any necessary software; and perform the actual data migration. A premigration data validation test is recommended, in addition to post-migration validation testing. These tests confirm that the data is in the same state after the migration as it was before. Clearly, the most important part of this stage is the migration itself. As outlined above, software technology can simplify this process by enhancing the speed of migration, by minimizing or eliminating application downtime, and/or by enabling migration during regular business hours, helping the organization to get back to business as quickly as possible.

Validate:

After the migration has been completed, the IT organization should compile migration statistics and prepare a report to highlight what worked, what didn't work and lessons learned. The report should be shared with all members of the migration team. These types of reports are critical in building a repeatable and consistent process through continuous process improvement—building on what worked and fixing or changing what didn't work. Further, documenting the migration process can help train staff, and simplify or streamline the next migration, reducing both expense and risk.

Conclusion:

Data migration is a routine part of IT operations in today's business environment. Even so, it often causes major disruptions as a result of downtime or application performance problems, and it can severely impact budgets. To prevent these problems, organizations need a consistent and reliable methodology that enables them to plan, design, migrate and validate the migration. Further, they need migration software that supports their specific migration requirements, including operating systems, storage platforms and performance. In addition, migration products that maintain continuous data availability during the migration without affecting performance are desirable.

Web References:

http://en.wikipedia.org/wiki/Data_migration

services.seagate.com/srs1-800-475-0143

http://www.infotechnet.org/ntca/DataMigration.htm

http://www-935.ibm.com/services/us/gts/pdf/softek-best-practices-data-migration.pdf

http://www.databaseanswers.org/data_migration/general_migration_approach.htm

http://www-935.ibm.com/services/us/gts/pdf/softek-best-practices-data-migration.pdf

IT 047

Honeypots for network security-How to track attackers activity

Miss. Sheetal Umbarkar

Asst. Prof. (Comp. Mgmt.),

ASM's IBMR, Chinchwad,

Pune, Maharashra, 411019,

Email: sheetalumbarkar@asmedu.org

Mobile No: 8087470228

Mrs. Haridini Bhagwat

Asst. Prof. (Comp. Mgmt.),

ASM's IBMR, Chinchwad,

Pune, Maharashra, 411019,

Email: haridini_dsrf@asmedu.org

Mobile No: 9405609554

ABSTRACT:

Honeypot is an exciting new technology with enormous potential for the security community.It is resource which is intended to be attacked and compromised to gain more information about the attacker and his attack techniques.

They are a highly flexible tool that comes in many shapes and sizes. This paper deals with understanding what a honeypot actually is ,and how it works. There are different varieties of honeypots. Based on their category they have different applications. This paper gives an insight into the use of honeypots in productive as well as educative environments.

This paper also discusses the advantages and disadvantages of honeypots , and what the future hold in store for them.

Keywords: Honeypot, Attack Techniques, Specter, Honeyd, and KFSensor, Symantec Decoy Server and Honeynets

INTRODUCTION:

The Internet is growing fast and doubling its number of websites every 53 days and the number of people using the internet is also growing. Hence, global communication is getting more important every day. At the same time, computer crimes are also increasing. Countermeasures are developed to detect or prevent attacks - most of these measures are based on known facts, known attack patterns. Countermeasures such as firewalls and network intrusion detection systems are based on prevention, detection and reaction mechanism; but is there enough information about the enemy?

As in the military, it is important to know, who the enemy is, what kind of strategy he uses, what tools he utilizes and what he is aiming for. Gathering this kind of information is not easy but important. By knowing attack strategies, countermeasure scan be improved and vulnerabilities can be fixed. To gather as much information as possible is one main goal of a honeypot. Generally, such information gathering should be done silently, without alarming an attacker. All the gathered information leads to an advantage on the defending side and can therefore be used on productive systems to prevent attacks.

A honeypot is primarily an instrument for information gathering and learning. Its primary purpose is not to be an ambush for the blackhat community to catch them in action and to press charges against them. The focus lies on a silent collection of as much information as possible about their attack patterns, used programs, purpose of attack and the blackhat community itself. All this information is used to learn more about the blackhat proceedings and motives, as well as their technical knowledge and abilities. This is just a primary purpose of a honeypot.

There are a lot of other possibilities for a honeypot - divert hackers from productive systems or catch a hacker while conducting an attack are just two possible examples. They are not the perfect solution for solving or preventing computer crimes.

Honeypots are hard to maintain and they need operators with good knowledge about operating systems and network security. In the right hands, a honeypot can be an effective tool for information gathering. In the wrong, unexperienced hands, a honeypot can become another infiltrated machine and an instrument for the blackhat community.

This paper will present the basic concepts behind honeypots and also the legal aspects of honeypots.

HONEYPOT BASICS:

Honeypots are an exciting new technology with enormous potential for the security community. The concepts were first introduced by several icons in computer security, specifically Cliff Stoll in the book The Cuckooâ„¢s Egg , and Bill Cheswick's paper "An Evening with Berferd. Since then, honeypots have continued to evolve, developing into the powerful security tools they are today.

Honeypots are neither like Firewalls that are used to limit or control the traffic coming into the network and to deter attacks neither is it like IDS (Intrusion Detection Systems) which is used to detect attacks. However it can be used along with these. Honeypots does not solve a specific problem as such, it can be used to deter attacks, to detect attacks, to gather information, to act as an early warning or indication systems etc. They can do everything from detecting encrypted attacks in IPv6 networks to capturing the latest in on-line credit card fraud. It is this flexibility that gives honeypots their true power. It is also this flexibility that can make them challenging to define and understand.

The basic definition of honeypots is:

A honeypot is an information system resource whose value lies in unauthorized or illicit use of that resource.

The main aim of the honeypot is to lure the hackers or attacker so as to capture their activities. This information proves to be very useful since information can be used to study the vulnerabilities of the system or to study latest techniques used by attackers etc. For this the honeypot will contain enough information (not necessarily real) so that the attackers get tempted. (Hence the name Honeypot â€œ a sweet temptation for attackers)Their value lies in the bad guys interacting with them. Conceptually almost all honeypots work they same. They are a resource that has no authorized activity, they do not have any production value.

Theoretically, a honeypot should see no traffic because it has no legitimate activity. This means any interaction with a honeypot is most likely unauthorized or malicious activity. Any connection attempts to a honeypot are most likely a probe, attack, or compromise. While this concept sounds very simple (and it is), it is this very simplicity that give honeypots their tremendous advantages (and disadvantages).

TYPES OF HONEYPOTS:

Honeypots come in many shapes and sizes, making them difficult to get a grasp of. To better understand honeypots and all the different types, they are broken down into two general categories, low-interaction and high-interaction honeypots. These categories helps to understand what type of honeypot one is dealing with, its strengths, and weaknesses. Interaction defines the level of activity a honeypot allows an attacker.

Low-interaction honeypots have limited interaction, they normally work by emulating services and operating systems. Attacker activity is limited to the level of emulation by the honeypot. For example, an emulated FTP service listening on port 21 may just emulate a FTP login, or it may support a variety of additional FTP commands.

The advantages of a low-interaction honeypot is their simplicity. These honeypots tend to be easier to deploy and maintain, with minimal risk. Usually they involve installing software, selecting the operating systems and services you want to emulate and monitor, and letting the honeypot go from there. This plug and play approach makes deploying them very easy for most organizations. Also, the emulated services mitigate risk by containing the attacker's activity, the attacker never has access to an operating system to attack or harm others.

The main disadvantages with low interaction honeypots is that they log only limited information and are designed to capture known activity. The emulated services can only do so much. Also, its easier for an attacker to detect a low-interaction honeypot, no matter how good the emulation is, skilled attacker can eventually detect their presence. Examples of low-interaction honeypots include Specter, Honeyd, and KFSensor.

High-interaction honeypots are different, they are usually complex solutions as they involve real operating systems and applications. Nothing is emulated, the attackers are given the real thing. If one wants a Linux honeypot running an FTP server, they build a real Linux system running a real FTP server. The advantages with such a solution are two fold. First, extensive amounts of information are captured. By giving attackers real systems to interact with, one can learn the full extent of the attackers behavior, everything from new rootkits to international IRC sessions.

The second advantage is high-interaction honeypots make no assumptions on how an attacker will behave. Instead, they provide an open environment that captures all activity. This allows high-interaction solutions to learn behavior one otherwise would not expect. An excellent example of this is how a Honeynet captured encoded back door commands on a non-standard IP protocol . However, this also increases the risk of the honeypot as attackers

can use these real operating system to attack non-honeypot systems. As result, additional technologies have to be implemented that prevent the attacker from harming other non-honeypot systems. In general, high-interaction honeypots can do everything low-interaction honeypots can do and much more. However, they can be more complex to deploy and maintain.

Examples of high-interaction honeypots include Symantec Decoy Server and Honeynets.

Low-interaction: Solution emulates operating systems and services.

High-interaction: No emulation, real OS and services are provided.Easy to install and deploy.

Captures limited amounts of information. Minimal risk, as the emulated services controls attackers . Can capture far more information Can be complex to install or deploy Increased risk, as attackers are provided real OS to interact with.

Some people also classify honeypots as low,mid and high interaction honeypots; where mid-interaction honeypots are those with their interaction level between that of low and high interaction honeypots.

A few examples of honeypots and their varieties are: BackOfficer Friendly BOF (as it is commonly called) is a very simple but highly useful honeypot developed by Marcus Ranum and crew at NFR. It is an excellent example of a low interaction honeypot.

It is a great way to introduce a beginner to the concepts and value of honeypots. BOF is a program that runs on most Window based operating system. All it can do is emulate some basic services, such as http, ftp, telnet, mail, or BackOrrifice. Whenever some attempts to connect to one of the ports BOF is listening to, it will then log the attempt. BOF also has the option of "faking replies", which gives the attacker something to connect to. This way one can log http attacks, telnet brute force logins, or a variety of other activity (Screenshot). The value in BOF is in detection, similar to a burglar alarm. It can monitor only a limited number of ports, but these ports often represent the most commonly scanned and targeted services.

Specter :

Specter is a commercial product and it is another 'low interaction' production honeypot. It is similar to BOF in that it emulates services, but it can emulate a far greater range of services and functionality. In addition, not only can it emulate services, but emulate a variety of operating systems.

Similar to BOF, it is easy to implement and low risk. Specter works by installing on a Windows system. The risk is reduced as there is no real operating system for the attacker to interact with. For example, Specter can emulate a web server or telnet server of the any operating system.

When an attacker connects, it is then prompted with an http header or login banner. The attacker can then attempt to gather web pages or login to the system. This activity is captured and recorded by Specter, however there is little else the attacker can do. There is no real application for the attacker to interact with, instead just some limited, emulated functionality. Specters value lies in detection. It can quickly and easily determine who is looking for what. As a honeypot, it reduces both false positives and false negatives, simplifying the detection process. Specter also supports a variety of alerting and logging mechanisms. You can see an example of this functionality in a screen shot of Specter.

One of the unique features of Specter is that it also allows for information gathering, or the automated ability to gather more information about the attacker. Some of this information gathering is relatively passive, such as Whois or DNS lookups. However, some of this research is active, such as port scanning the attacker.

Homemade Honeypots

Another common honeypot is homemade. These honeypots tend to be low interaction. Their purpose is usually to capture specific activity, such as Worms or scanning activity. These can be used as production or research honeypots, depending on their purpose. Once again, there is not much for the attacker to interact with, however the risk is reduced because there is less damage the attacker can do. One common example is creating a service that listens on port 80 (http) capturing all traffic to and from the port. This is commonly done to capture Worm attacks Homemade honeypots can be modified to do (and emulate) much more, requiring a higher level of involvement, and incurring a higher level of risk. For example, FreeBSD has a jail functionality, allowing an administrator to create a controlled environment within the operating system. The attacker can then interact with this controlled environment. The value here is the more the attacker can do, the more can be potentially learned. However, care must be taken, as the more functionality the attacker can interact with, the more can go wrong, with the honeypot potentially compromised.

Honeyd:

Created by Niels Provos, Honeyd is an extremely powerful, OpenSource honeypot. Designed to run on Unix systems, it can emulate over 400 different operating systems and thousands of different computers, all at the same time. Honeyd introduces some exciting new features. First, not only does it emulate operating systems at the application level, like Specter, but it also emulates operating systems at the IP stack level. This means when someone Nmaps the honeypot, both the service and IP stack behave as the emulated operating system. Currently no other honeypot has this capability (CyberCop Sting did have this capability, but is no longer available). Second, Honeyd can emulate hundreds if not thousands of different computers all at the same time. While most honeypots can only emulate one computer at any point in time, Honeyd can assume the identity of thousands of different IP addresses. Third, as an OpenSource solution, not only is it free to use, but it will expotentially grow as members of the security community develop and contribute code.

Honeyd is primarily used for detecting attacks. It works by monitoring IP addresses that are unused, that have no system assigned to them. Whenever an attacker attempts to probe or attack an non-existant system, Honeyd, through Arp spoofing, assumes the IP address of the victim and then interacts with the attacker through emulated services. These emulates services are nothing more then scripts that react to predetermined actions. For example, a script can be developed to behave like a Telnet service for a Cisco router, with the Cisco IOS login interface. Honeyd's emulated services are also Open Source, so anyone can develop and use their own. The scripts can be written in almost any language, such as shell or Perl. Once connected, the attacker believes they are interacting with a real system. Not only can Honeyd dynamically interact with attackers, but it can detect activity on any port. Most low interaction honeypots are limited to detecting attacks only on the ports that have emulated services listening on. Honeyd is different, it detects and logs connections made to any port, regardless if there is a service listening. The combined capabilities of assuming the identity of non-existant systems, and the ability to detect activity on any port, gives Honeyd incredible value as a tool to detect unauthorized activity. I highly encourage people to check it out, and if possible to contribute new emulated services.

Mantrap :

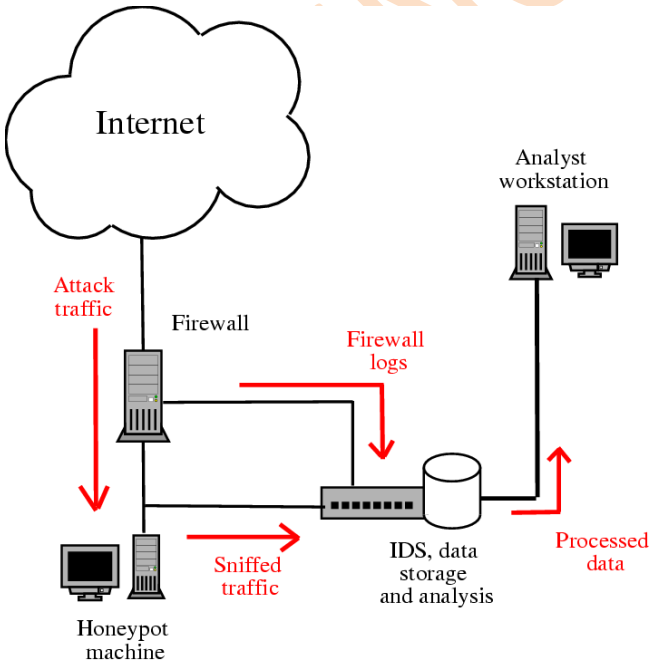
Produced by Recourse, Mantrap is a commercial honeypot. Instead of emulating services, Mantrap creates up to four sub-systems, often called 'jails'. These 'jails' are logically discrete operating systems separated from a master operating system (see Diagram.) Security administrators can modify these jails just as they normally would with any operating system, to include

installing applications of their choice, such as an Oracle database or Apache web server. This makes the honeypot far more flexible, as it can do much more. The attacker has a full operating system to interact with, and a variety of applications to attack. All of this activity is then captured and recorded. Not only can we detect port scans and telnet logins, but we can capture rootkits, application level attacks, IRC chat session, and a variety of other threats. However, just as far more can be learned, so can more go wrong. Once compromised, the attacker can used that fully functional operating system to attack others. Care must be taken to mitigate this risk. As such, it can be categorized this as a mid-high level of interaction. Also, these honeypots can be used as either a production honeypot (used both in detection and reaction) or a research honeypot to learn more about threats. There are limitations to this solution. The biggest one is that we are limited to only what the vendor supplies us. Currently, Mantrap only exists on Solaris operating system.

Honeynets:

Honeynets represent the extreme of research honeypots. They are high interaction honeypots, one can learn a great deal, however they also have the highest level of risk.

Fig: A honeynet:



Their primary value lies in research, gaining information on threats that exist in the Internet community today. A Honeynet is a network of production systems. Unlike many of the honeypots discussed so far, nothing is emulated. Little or no modifications are made to the honeypots. The idea is to have an architecture that creates a highly controlled network, one where all activity is controlled and captured. Within this network we place our intended victims, real computers running real applications. The bad guys find, attack, and break into these systems on their own initiative. When they do, they do not realize they are within a Honeynet. This gives the attackers a full range of systems, applications, and functionality to attack. All of their activity, from encrypted SSH sessions to emails and files uploads, are captured without them knowing it. This is done by inserting kernel modules on the victim systems that capture all of the attacker's actions. From this we can learn a great deal, not only their tools and tactics, but their methods of communication, group organization, and motives. However, with this capability comes a great deal of risk.

A variety of measures must be taken to ensure that once compromised, a Honeynet cannot be used to attack others. Honeynets do this using a Honeywall gateway. This gateway allows inbound traffic to the victim systems, but controls the outbound traffic using intrusion prevention technologies. This gives the attacker the flexibility to interact with the victim systems, but prevents the attacker from harming other non-Honeynet computers. Honeynets are primarily research honeypots. They could be used as production honeypots, specifically for detection or reaction, however it is most likely not worth the time and effort
We have reviewed six different types of honeypots. No one honeypot is better than the other, each one has its advantages and disadvantages, it all depends on what is to be achieved. To more easily define the capabilities of honeypots, we have categorized them based on their level of interaction. The greater interaction an attacker has, the more we can learn, but the greater the risk. For example, BOF and Specter represent low interactions honeypots. They are easy to deploy and have minimal risk. However, they are limited to emulating specific services and operating systems, used primarily for detection. Mantrap and Honeynets represent mid-to-high interaction honeypots. They can give far greater depth of information, however more work and greater risk is involved Sometimes, honeypots are also classified as Hardware based and Software based honeypots.

Hardware-based honeypots are servers, switches or routers that have been partially disabled and made attractive with commonly known misconfigurations. They sit on the internal network, serving no purpose but to look real to outsiders. The operating system of each box, however, has been

subtly disabled with tweaks that prevent hackers from really taking it over or using it to launch new attacks on other servers.

Software emulation honeypots, on the other hand, are elaborate deception programs that mimic real Linux or other servers and can run on machines as low-power as a 233-MHz PC. Since an intruder is just dancing with a software decoy, at no time does he come close to actually seizing control of the hardware, no matter what the fake prompts seem to indicate. Even if the hacker figures out that it's a software honeypot, the box on which it's running should be so secure or isolated that he couldn't do anything but leave anyway.Software emulation might be more useful for corporate environments where business secrets are being safeguarded.

VALUE OF HONEYPOTS:

Now that we have understanding of two general categories of honeypots, we can focus on their value. Specifically, how we can use honeypots. Once again, we have two general categories, honeypots can be used for production purposes or research. When used for production purposes, honeypots are protecting an organization. This would include preventing, detecting, or helping organizations respond to an attack. When used for research purposes, honeypots are being used to collect information. This information has different value to different organizations. Some may want to be studying trends in attacker activity, while others are interested in early warning and prediction, or law enforcement. In general, low-interaction honeypots are often used for production purposes, while high-interaction honeypots are used for research purposes. However, either type of honeypot can be used for either purpose. When used for production purposes, honeypots can protect organizations in one of three ways; prevention, detection, and response. We will take a more in-depth look at how a honeypot can work in all three.

1. Prevention : Honeypots can help prevent attacks in several ways. The first is against automated attacks, such as worms or auto-rooters. These attacks are based on tools that randomly scan entire networks looking for vulnerable systems. If vulnerable systems are found, these automated tools will then attack and take over the system (with worms self-replicating, copying themselves to the victim). One way that honeypots can help defend against such attacks is slowing their scanning down, potentially even stopping them. Called sticky honeypots, these solutions monitor unused IP space. When probed by such scanning activity, these honeypots interact with and slow the attacker down. They do this using a variety of TCP tricks, such as a Windows size of zero, putting the attacker into a holding pattern. This is excellent for slowing down or preventing the spread of a worm that has penetrated the

internal organization. One such example of a sticky honeypot is LaBrea Tarpit. Sticky honeypots are most often low-interaction solutions (one can almost call them 'no-interaction solutions', as they slow the attacker down to a crawl).

Honeypots can also be used to protect the organization from human attackers. The concept is deception or deterrence. The idea is to confuse an attacker, to make him waste his time and resources interacting with honeypots. Meanwhile, the organization being attacked would detect the attacker's activity and have the time to respond and stop the attacker. This can be even taken one step farther. If an attacker knows an organization is using honeypots, but does not know which systems are honeypots and which systems are legitimate computers, they may be concerned about being caught by honeypots and decided not to attack your organizations. Thus the honeypot deters the attacker. An example of a honeypot designed to do this is Deception Toolkit, a low-interaction honeypot.

2. Detection : The second way honeypots can help protect an organization is through detection. Detection is critical, its purpose is to identify a failure or breakdown in prevention. Regardless of how secure an organization is, there will always be failures, if for no other reasons then humans are involved in the process. By detecting an attacker, you can quickly react to them, stopping or mitigating the damage they do. Traditionally, detection has proven extremely difficult to do. Technologies such as IDS sensors and systems logs have proved ineffective for several reasons. They generate far too much data, large percentage of false positives (i.e. alerts that were generated when the sensor recognized the configured signature of an "attack", but in reality was just valid traffic), inability to detect new attacks, and the inability to work in encrypted or IPv6 environments. Honeypots excel at detection, addressing many of these problems of traditional detection. Since honeypots have no production activity, all connections to and from the honeypot are suspect by nature. By definition, anytime a connection is made to the honeypot, this is most likely an unauthorized probe, scan, or attack. Anytime the honeypot initiates a connection, this most likely means the system was successfully compromised. This helps reduce both false positives and false negatives greatly simplifying the detection process by capturing small data sets of high value, it also captures unknown attacks such as new exploits or polymorphic shellcode, and works in encrypted and IPv6 environments. In general, low-interaction honeypots make the best solutions for detection. They are easier to deploy and maintain then high-interaction honeypots and have reduced risk.

3. Response : The third and final way a honeypot can help protect an organization is in reponse. Once an organization has detected a failure, how do they respond? This can often be one of the greatest challenges an organization faces. There is often little information on who the attacker is, how they got in, or how much damage they have done. In these situations detailed information on the attacker's activity are critical. There are two problems compounding incidence response. First, often the very systems compromised cannot be taken offline to analyze. Production systems, such as an organization's mail server, are so critical that even though its been hacked, security professionals may not be able to take the system down and do a proper forensic analysis. Instead, they are limited to analyze the live system while still providing production services. This cripples the ability to analyze what happened, how much damage the attacker has done, and even if the attacker has broken into other systems. The other problem is even if the system is pulled offline, there is so much data pollution it can be very difficult to determine what the bad guy did. By data pollution, I mean there has been so much activity (user's logging in, mail accounts read, files written to databases, etc) it can be difficult to determine what is normal day-to-day activity, and what is the attacker.

Honeypots can help address both problems. Honeypots make an excellent incident resonse tool, as they can quickly and easily be taken offline for a full forensic analysis, without impacting day-to-day business operations. Also, the only activity a honeypot captures is unauthorized or malicious activity. This makes hacked honeypots much easier to analyze then hacked production systems, as any data you retrieve from a honeypot is most likely related to the attacker. The value honeypots provide here is quickly giving organizations the in-depth information they need to rapidly and effectively respond to an incident. In general, high-interaction honeypots make the best solution for response. To respond to an intruder, you need in-depth knowledge on what they did, how they broke in, and the tools they used. For that type of data you most likely need the capabilities of a high-interaction honeypot.

Up to this point we have been talking about how honeypots can be used to protect an organization. We will now talk about a different use for honeypots, research.

Honeypots are extremely powerful, not only can they be used to protect your organization, but they can be used to gain extensive information on threats, information few other technologies are capable of gathering. One of the greatest problems security professionals face is a lack of information or intelligence on cyber threats. How can we defend against an enemy when we don't even know who that enemy is? For centuries military organizations have depended on information to better understand who their enemy is and how to

defend against them. Why should information security be any different?

Research honeypots address this by collecting information on threats. This information can then be used for a variety of purposes, including trend analysis, identifying new tools or methods, identifying attackers and their communities, early warning and prediction, or motivations. One of the most well known examples of using honeypots for research is the work done by the Honeynet Project, an all volunteer, non-profit security research organization. All of the data they collect is with Honeynet distributed around the world. As threats are constantly changing, this information is proving more and more critical.

IMPLEMENTATION:

Honeypot Location:

A honeypot does not need a certain surrounding environment as it is a standard server with no special needs.A honeypot can be placed anywhere a server could be placed. But certainly, some places are better for certain approaches as others.

A honeypot can be used on the Internet as well as the intranet, based on the needed service. Placing a honeypot on the intranet can be useful if the detection of some bad guys inside a private network is wished. It is especially important to set the internal thrust for a honeypot as low as possible as this system could be compromised, probably without immediate knowledge.

If the main concern is the Internet, a honeypot can be placed at two locations:

1. In front of the firewall (Internet) DMZ
2. Behind the firewall (intranet)

Each approach has its advantages as well as disadvantages. Sometimes it is even impossible to choose freely as placing a server in front of a firewall is simply not possible or not wished.

By placing the honeypot in front of a firewall , the risk for the internal network does not increase. The danger of having a compromised system behind the firewall is eliminated. A honeypot will attract and generate a lot of unwished traffic like portscans or attack patterns. By placing a honeypot outside the firewall, such events do not get logged by the firewall and an internal IDS system will not generate alerts. Otherwise, a lot of alerts would be generated on the firewall or IDS.Probably the biggest advantage is that the

firewall or IDS, as well as any other resources, have not to be adjusted as the honeypot is outside the firewall and viewed as any other machine on the external network. The disadvantage of placing a honeypot in front of the firewall is that internal attackers cannot be located or trapped that easy, especially if the firewall limits outbound traffic and therefore limits the traffic to the honeypot.

Placing a honeypot inside a DMZ seems a good solution as long as the other systems inside the DMZ can be secured against the honeypot. Most DMZs are not fully accessible as only needed services are allowed to pass the firewall. In such a case,placing the honeypot in front of the firewall should be favored as opening all corresponding ports on the firewall is too time consuming and risky.

A honeypot behind a firewall can introduce new security risks to the internal network, especially if the internal network is not secured against the honeypot through additional firewalls. This could be a special problem if the IPâ„¢s are used for authentication. It is important to distinguish between a setup where the firewall enables access to the honeypot or where access from the Internet is denied. By placing the honeypot behind a firewall, it is inevitable to adjust the firewall rules if access from the Internet should be permitted. The biggest problem arises as soon as the internal honeypot is compromised by an external attacker. He gains the possibility to access the internal network through the honeypot. This traffic will be unstopped by the firewall as it is regarded as traffic to the honeypot only, which in turn is granted. Securing an internal honeypot is therefore mandatory, especially if it is a high-involvement honeypot. With an internal honeypot it is also possible to detect a misconfigured firewall which forwards unwanted traffic from the Internet to the internal network. The main reason for placing a honeypot behind a firewall could be to detect internal attackers. The best solution would be to run a honeypot in its own DMZ, therefore with a preliminary firewall. The firewall could be connected directly to the Internet or intranet, depending on the goal. This attempt enables tight control as well as a flexible environment with maximal security.

How does a Honeypot Gather Information:

Obviously a honeypot must capture data in an area that is not accessible to an attacker. Data capture happens on a number of levels. Firewall Logs

A Packet Sniffer (or similar IDS sensor)

The IDS should be configured to passively monitor network traffic (for an added level of invisibility, one might set the system up to have no IP address or, in some instances, the sniffer could be configured to completely lack an IP

stack). This will capture all cleartext communication, and can read keystrokes.

1.Local and Remote Logs: These should be set up just as it would on any other system, and will possibly be disabled, deleted, or modified by an experienced hacker, but plenty of useful information will still be available from all the previous capture methods.

2. Remotely Forwarded Logs: Will capture data on a remote log and then instantly forward the data to a system even further out of the range of the attacker, so that the attacker cannot be warned that all his activities are watched or try to modify the captured data.

Limiting Outbound Attacks:

To protect oneself from any sort of third party liabilities, an individual deploying a honeypot will likely want some kind of safeguard. Firewalls can be configured to let an unlimited number of inbound connections, while limiting outbound connections to a specific number (be it 10 outbound connections, or 50). This method lacks flexibility, and could shut an attacker out at a critical point (in the middle of an IRC session, or before they have retrieved all of their tools). A more flexible option is as follows: a system configured as a layer 2 bridge (which will lack all TCP activity, thus being harder to detect). The system can be configured to monitor all activity and can utilize a signature database to distinguish a known attack from any non-aggressive activity (and instead of blocking the attack, it can simply add some data to the packet to render it ineffectual). It can also throttle bandwidth (to quench a DDoS attack). This is a very effective way to protect other systems; however, it will not block unknown or new attacks.

Putting the Honey into the Pot

An advanced honeypot is a fully functional OS, and therefore can be filled with financial information, e-mails with passwords for other honeypots, databases of fake customers anything that might motivate an attacker to compromise the system. An individual could set up a web server that explains that the law services of so and so and so and so from San Francisco are currently setting up their systems to do online consultation for big banks and other big businesses. A whole network of honeypots sits in a secure environment behind a firewall that an attacker would need to break through. The network might have loads of fake data and e-mail; a large playing field for an advanced hacker to wander through.

MERITS AND DEMERITS:

Merits: Honeypots have a large number of merits in its favour.

They are :

Small data sets of high value: Honeypots collect small amounts of information. Instead of logging a one GB of data a day, they can log only one MB of data a day. Instead of generating 10,000 alerts a day, they can generate only 10 alerts a day. Remember, honeypots only capture bad activity, any interaction with a honeypot is most likely unauthorized or malicious activity. As such, honeypots reduce 'noise' by collectin only small data sets, but information of high value, as it is only the bad guys. This means its much easier (and cheaper) to analyze the data a honeypot collects and derive value from it.

New tools and tactics: Honeypots are designed to capture anything thrown at them, including tools or tactics never seen before.
Minimal resources: Honeypots require minimal resources, they only capture bad activity. This means an old Pentium computer with 128MB of RAM can easily handle an entire class B network sitting off an OC-12 network.

Encryption or IPv6: Unlike most security technologies (such as IDS systems) honeypots work fine in encrypted or IPv6 environments. It does not matter what the bad guys throw at a honeypot, the honeypot will detect and capture it.

Information: Honeypots can collect in-depth information that few, if any other technologies can match.

Simplicty: Finally, honeypots are conceptually very simple. There are no fancy algorithms to develop, state tables to maintain, or signatures to update. The simpler a technology, the less likely there will be mistakes or misconfigurations.

Demerits:

Like any technology, honeyopts also have their weaknesses. It is because of this they do not replace any current technology, but work with existing technologies.

Limited view: Honeypots can only track and capture activity that directly interacts with them. Honeypots will not capture attacks against other systems, unless the attacker or threat interacts with the honeypots also.

Risk: All security technologies have risk. Firewalls have risk of being penetrated, encryption has the risk of being broken, IDS sensors have the risk of failing to detect attacks. Honeypots are no different, they have risk also. Specifically, honeypots have the risk of being taken over by the bad guy and being used to harm other systems. This risk various for different honeypots. Depending on the type of honeypot, it can have no more risk then an IDS sensor, while some honeypots have a great deal of risk.

LEGAL ISSUES:

In the past there has been some confusion on what are the legal issues with honeypots. There are several reasons for this. First, honeypots are relatively new. Second, honeypots come in many different shapes and sizes and accomplish different goals. Based on the different uses of honeypots different legal issues apply. Last, there are no precedents for honeypots. There are no legal cases recorded on the issues. The law is developed through cases. Without cases directly on point, we are left trying to predict, based on cases in other contexts, how courts will treat honeypots. Until a judge gives a court order, we will really never know.

With honeypots, there are three main issues that are commonly discussed: entrapment, privacy, and liability.

Liability: You can potentially be held liable if your honepyot is used to attack or harm other systems or organizations. This risk is the greatest with Research honeypots.

Privacy: Honeypots can capture extensive amounts of information about attackers, which can potentially violate their privacy. Once again, this risk is primarily with Research honeypots. However in case of honeypot there is exemption. It means that security technologies can collect information on people (and attackers), as long as that technology is being used to protect or secure your environment. In other words, these technologies are now exempt from privacy restrictions. For example, an IDS sensor that is used for detection and captures network activity is doing so to detect (and thus enable organizations to respond to) unauthorized activity. Such a technology is most

likely not considered a violation of privacy.

Entrapment: For some odd reason, many people are concerned with the issue of entrapment. Entrapment, by definition is "a law-enforcement officer's or government agent's inducement of a person to commit a crime, by means of fraud or undue persuasion, in an attempt to later bring a criminal prosecution against that person." Think about it, entrapment is when you coerce or induce someone to do something they would not normally do. Honeypots do not induce anyone. Attackers find and break into honeypots on their own initiative. People often question the idea of creating targets of high value, for example honeypots that are ecommerce sites or advertised as having government secrets. Even then, such honeypots are most likely not a form of entrapment as you are not coercing them into breaking into the honeypot. The bad guy has already decided to commit unauthorized activity, one is merely providing a different target for the blackhat to attack. Therefore, in most cases involving honeypots, entrapment is not an issue.

FUTURE OF HONEYPOTS:

Mr. Lance spitzner who has played a major role in the development of honeypots has made certain predictions about the future of honeypots. They are as follows:

Government projects: Currently honeypots are mainly used by organizations, to detect intruders within the organization as well as against external threats and to protect the organization. In future, honeypots will play a major role in the government projects, especially by the military, to gain information about the enemy, and those trying to get the government secrets.

Ease of use: In future honeypots will most probably appear in prepackaged solutions, which will be easier to administer and maintain. People will be able to install and develop honeypots at home and without difficulty.

Closer integration: Currently honeypots are used along with other technologies such as firewall, tripwire, IDS etc. As technologies are developing, in future honeypots will be used in closer integration with them.

For example honeypots are being developed for WI-FI or wireless computers. However the development is still under research.

Â¢ Specific purpose: Already certain features such as honeytokens are under development to target honeypots only for a specific purpose. Eg: catching only those attempting credit card fraud etc.

Honeypots will be used widely for expanding research applications in future.

CONCLUSION:

This paper has given an in depth knowledge about honeypots and their contributions to the security community. A honeypot is just a tool. How one uses this tool is upto them.

Honeypots are in their infancy and new ideas and technologies will surface in the next time. At the same time as honeypots are getting more advanced, hackers will also develop methods to detect such systems. A regular arms race could start between the good guys and the blackhat community. Lets hope that such a technology will be used to restore the peace and prosperity of the world and not to give the world a devastating end.

REFERENCES:

- Spitzner, Lance., Honeypots Tracking Hackers. Addison-Wesley: Boston,2002
- Spitzner, Lance.,The value of Honeypots, Part Two:Honeypot Solutions and legal Issues , 10 Nov.2002, http://online.securityfocus.com/infocus/1498>S
- Spitzner, Lance.,Know Your Enemy: Honeynets. 18 Sep. 2002. http://project.honeynet.org/papers/honeynet/
- Honeypots-Turn the table on hackers June 30,2003

www.itmanagement.earthweb.com/secu/article.php/143 6291

IT 048

Upcoming Os in market"Androis" Android

Submitted by: Prof : Swati Jadhav.

swatijadhav@asmedu.org Mob:- 9822251608.

(ASM's IBMR Chinchwad Pune.)

Abstract

Android is a Open Source Project originated by a group of companies known as the Open Handset Alliance, led by Google. Android is a Mobile Operating System. Android Operating System began with the release of the Android 1.0 beta in November 2007. In April 2009, each Android version has been developed under a codename based on a dessert item. The versions have released in alphabetical order: Cupcake, Donut, Eclair, Froyo, Gingerbread, Honeycomb and Ice Cream Sandwich. it has number of updates to its base operating system. Android breaks down the barriers to building new and innovative applications.Android provides access to a wide range of useful libraries and tools. Android platform will be made available under the Apache free-software and open-source license.

Keywords: Open Handset Alliance(OHA), Architecture,Platform,Compatibility, Features,Applications,Advantages,Limitations & New Versions.

Introduction: It is widely acknowledged today that new technologies in particular access to the Internet, tend to modify communication between the different players in the world .Search Engine "Google" never forgets anything, it comes up with pretty interesting ideas related to OS .Operating System is a set of Programs which manages hardware resources & provide services for application softwares. it is important for computer system as well now with mobile devices i.e "Android" the worlds most popular mobile platform. Android is a software stack for mobile devices that includes an operating system, middleware and key applications. Android is based on the Linux operating system and developed by Google and the Open Handset Alliance. It allows developers to write managed code in a Java-like languages.Android is a polished & professional home page for Google's mobile Platform. But beneath that elegant exterior lies a treasure trove of fascinating information. No shocking secrets are revealed, we will be getting a rare behind-the-scenes glimpse at the history of the Android and the evolution of its famous domain.

History: Android Operating System began with the release of the Android 1.0 beta in November 2007. It was released on 5th November 2007.There has been number of updates to its base operating system since its original release. Android 1.0, the first commercial version of the software, was released on 23 September 2008.The updates typically fix bugs and add new features. On 9 February 2009, the Android 1.1 update was released, initially for the T-Mobile only. The update resolved bugs ,changed the API and added a number of other features.

On 30 April 2009, the Android 1.5 update, dubbed Cupcake, was released, based on Linux kernel.The update included several new features and UI amendments.

Later on New Versions such as Donut, Eclair, Froyo, Gingerbread, Honeycomb and Ice Cream Sandwich were launched with the updates.

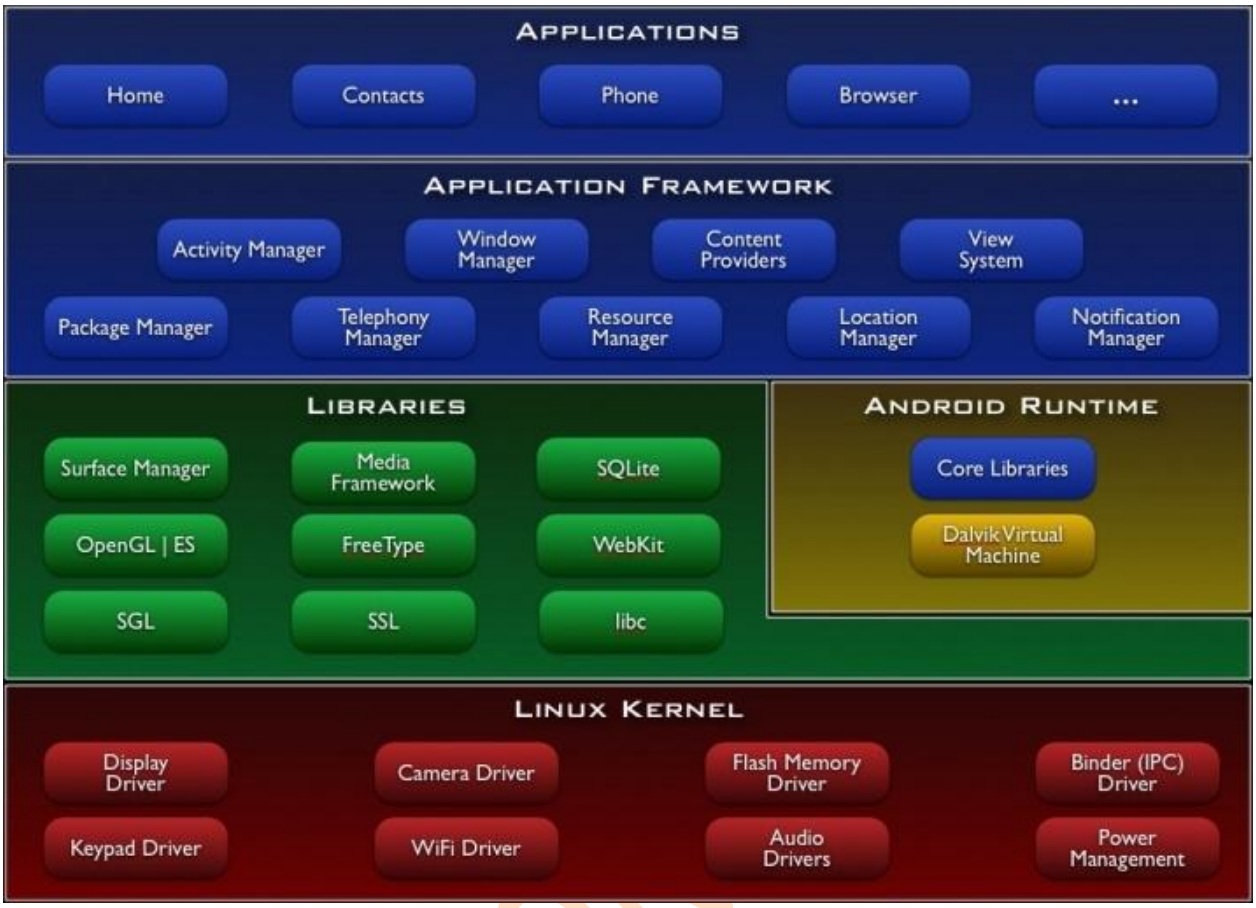
Currently Ice cream sandwich is lauched in market with new faetures,Functionality & User Interface.

Open Handset Alliance(OHA):

On 5 November 2007, the Open Handset Alliance, a consortium of several companies which include Google, HTC, Intel, Motorola, Qualcomm, T-Mobile, Sprint Nextel and NVIDIA, was unveiled with the goal to develop open standards for mobile devices. Along with the formation of the Open Handset Alliance, the OHA also unveiled their first product, Android, an open source mobile device platform based on the Linux operating system.The Open Handset Alliance (OHA) is a consortium of 84 firms to develop open standards for mobile devices.it was from the ground-up to enable developers to create compelling mobile applications that take full advantage of all a handset has to offer. It was built to be truly open. For example, an application can call upon any of the phone's core functionality such as making calls, sending text messages, or using the camera, allowing developers to create richer and more cohesive experiences for users. Android is also open source, which means developers can customize the OS for different phones and applications. This is why different phones may have different looking graphical interfaces and features even though they are running the same OS

Architecture:

The following diagram shows the major components of the Android operating system.Consist of six software layers :Application,Application Framework,Runtime,Libraries,Android Runtime & Linux Kernel. Each section is described in more detail below.



Applications

Android will ship with a set of core applications including an email client, SMS program, calendar, maps, browser, contacts, and others. All applications are written using the Java programming language.

Application Framework

By providing an open development platform, Android offers developers the ability to build extremely rich and innovative applications. Developers are free to take advantage of the device hardware, access location information, run background services, set alarms, add notifications to the status bar, and much, much more.

Developers have full access to the same framework APIs used by the core applications. The application architecture is designed to simplify the reuse of components; any application can publish its capabilities and any other application may then make use of those capabilities (subject to security constraints enforced by the framework). This same mechanism allows components to be replaced by the user.

Underlying all applications is a set of services and systems, including:

- A rich and extensible set of view that can be used to build an application, including lists, grids, text boxes, buttons, and even an embeddable web browser
- Content Providers that enable applications to access data from other applications (such as Contacts), or to share their own data
- A Resource Manager providing access to non-code resources such as localized strings, graphics, and layout files
- A Notification Manager that enables all applications to display custom alerts in the status bar
- An Activity Manager that manages the lifecycle of applications and provides a common navigation backstack.

Libraries: Android includes a set of C/C++ libraries used by various components of the Android system. These capabilities are exposed to developers through the Android application framework. Some of the core libraries are listed below:

- **System C library** - a BSD-derived implementation of the standard C system library (libc), tuned for embedded Linux-based devices
- **Media Libraries** - based on PacketVideo's OpenCORE; the libraries support playback and recording of many popular audio and video formats, as well as static image files, including MPEG4, H.264, MP3, AAC, AMR, JPG, and PNG
- **Surface Manager** - manages access to the display subsystem and seamlessly composites 2D and 3D graphic layers from multiple applications
- **LibWebCore** - a modern web browser engine which powers both the Android browser and an embeddable web view
- **SGL** - the underlying 2D graphics engine
- **3D libraries** - an implementation based on OpenGL ES 1.0 APIs; the libraries use either hardware 3D acceleration (where available) or the included, highly optimized 3D software rasterizer
- **FreeType** - bitmap and vector font rendering
- **SQLite** - a powerful and lightweight relational database engine available to all applications

Android Runtime: Android includes a set of core libraries that provides most of the functionality available in the core libraries of the Java programming language.Every Android application runs in its own process, with its own instance of the Dalvik virtual machine. Dalvik has been written so that a device can run multiple VMs efficiently. The Dalvik VM executes files in the Dalvik Executable (.dex) format which is optimized for minimal memory footprint. The VM is register-based, and runs classes compiled by a Java language compiler that have been transformed into the .dex format by the included "dx" tool.The Dalvik VM relies on the Linux kernel for underlying functionality such as threading and low-level memory management

Linux Kernel: Android relies on Linux version 2.6 for core system services such as security, memory management, process management, network stack,

and driver model. The kernel also acts as an abstraction layer between the hardware and the rest of the software stack.

Platform:

Android OS platform has a wide range of new features, as well as a refurbished user interface. Android is built on the open Linux Kernel.

Android is built on the open Linux Kernel.it is a project-management technique. We view and manage Android as a single, holistic software product, not a "distribution", specification, or collection of replaceable parts. Our intent is that device builders port Android to a device; they don't implement a specification or a distribution. Android relies on Linux version 2.6 for core system services such as security, memory management, process management, network stack, and driver model. The kernel also acts as an abstraction layer between the hardware and the rest of the software stack.Android includes a set of C/C++ libraries used by various components of the Android system. These capabilities are exposed to developers through the Android application framework.The robustness and scalability of the platform is secured by the community and open access to a central repository of updated code. Android includes almost the entirety of the applications-related software stack, less key technical pieces such as telephony protocol stacks, which are left to silicon vendors. Android bundles critical components such as a Linux kernel from Wind River, various optimized graphics engines, codecs, notification software, a clean room JVM (Java Virtual Machine)implementation, and the KHTML open source browser.

Compatibility:

Android's purpose is to establish an open platform for developers to build innovative mobile apps. Three key components work together to realize this platform.The Android SDK(Software Development Kit) provides built-in tools that Developers use to clearly state the device features their apps requires. And Android Market shows apps only to those devices that can properly run them.

Users want a customizable device.A mobile phone is a highly personal, always-on, always-present gateway to the Internet. We haven't met a user yet who didn't want to customize it by extending its functionality. That's why Android was designed as a robust platform for running after-market applications.

Developers outnumber us all. No device manufacturer can hope to write all the software that a person could conceivably need. We need third-party developers to write the apps users want, so the Android Open Source Project aims to make it as easy and open as possible for developers to build apps.

Everyone needs a common ecosystem. Every line of code developers write to work around a particular phone's bug is a line of code that didn't add a new

feature. The more compatible phones there are, the more apps there will be. By building a fully compatible Android device, you benefit from the huge pool of apps written for Android, while increasing the incentive for developers to build more of those apps.

Android compatibility is free, and it's easy. If you are building a mobile device, you can follow these steps to make sure your device is compatible with Android. For more details about the Android compatibility program in general

Building a compatible device is a three-step process:

1. Obtain the Android software source code that you port to your hardware.
2. Comply with Android Compatibility Definition Document (CDD). The CDD enumerates the software and hardware requirements of a compatible Android device.
3. Pass the Compatibility Test Suite (CTS). You can use the CTS (included in the Android source code) as an ongoing aid to compatibility during the development process.

Features:

DalviK Virtual Machine:it is extremely low memory based virtual machine which was designed especially for Android to run on embedded systems and work well in low power situations.it is also turned to the cpu attributes.it creates a special file format (.dex) that is created through build processing.Conversion between java classes and .dex format is done by included dx tool.

Integrated Browser : Google made a right choice on choosing WebKit as open source web browser. They added a two pass layout and frame flattening. Two pass layout loads a page without waiting for blocking elements, such as external CSS or external JavaScript and after a while renders again with all resources downloaded to the device. Frame flattening converts founded frames into single one and loads into the browser. These features increase speed and usability browsing the internet via mobile phone.

Optimized Graphics: Android has 2D graphics library and 3D graphics based on OpenGL ES 1.0, possibly we will see great applications like Google Earth and spectacular games like Second Life, which come on Linux version. At this moment, the shooting legendary 3D game Doom was presented using Android on the mobile phone.

Data Storage :SQLite is used for structured data storage .SQLite is a powerful and lightweight relational database engine available to all applications.

Connectivity :Android supports a wide variety of connectivity technologies including GSM, CDMA, Bluetooth, EDGE, EVDO, 3G and Wi-Fi.

Messaging:SMS, MMS, and XMPP are available forms of messaging including threaded text messaging.

Web Browser:The web browser available in Android is based on the open-source WebKit application framework. It includes LibWebCore which is a modern web browser engine which powers both the Android browser and an embeddable web view.

Java Virtual Machine :Software written in Java can be compiled into Dalvik bytecodes and executed in the Dalvik virtual machine, which is a specialized VM implementation designed for mobile device use, although not technically a standard Java Virtual Machine.

Media Support :Android will support advanced audio/video/still media formats such as MPEG-4, H.264, MP3, and AAC, AMR, JPEG, PNG, GIF. Additional Hardware Support :Android is fully capable of utilizing video/still cameras, touchscreens, GPS, compasses, accelerometers, and accelerated 3D graphics.

Development Environment :Includes a device emulator, tools for debugging, memory and performance profiling, a plugin for the Eclipse IDE. There are a number of hardware dependent features, for instance, a huge media and connections support, GPS, improved support for Camera and simply GSM telephony. A great work was done for the developers to start work with Android using device emulator, tools for debugging and plugin for Eclipse IDE.

Applications:

An Android application is composed of more than just code—it requires resources that are separate from the source code, such as images, audio files, and anything relating to the visual presentation of the application.

- Android 4.0 makes common actions more visible and lets users navigate with simple, intuitive gestures. Refined animations and feedback throughout the system make interactions engaging and interesting. An entirely new typeface optimized for high-resolution screens improves readability and brings a polished, modern feel to the user interface.
- Multitasking is a key strength of Android and it's made even easier and more visual on Android 4.0.
- The Android operating system is a multi-user Linux system in which each application is a different user.
- By default, the system assigns each application a unique Linux user ID (the ID is used only by the system and is unknown to the application). The system

sets permissions for all the files in an application so that only the user ID assigned to that application can access them.

- Each process has its own virtual machine (VM), so an application's code runs in isolation from other applications.
- By default, every application runs in its own Linux process. Android starts the process when any of the application's components need to be executed, then shuts down the process when it's no longer needed or when the system must recover memory for other applications.
- Android breaks down the barriers to building new and innovative applications. For example, a developer can combine information from the web with data on an individual's mobile phone — such as the user's contacts, calendar, or geographic location — to provide a more relevant user experience. With Android, a developer can build an application that enables users to view the location of their friends and be alerted when they are in the vicinity giving them a chance to connect.
- Fast & easy application development. Android provides access to a wide range of useful libraries and tools that can be used to build rich applications. For example, Android enables developers to obtain the location of the device, and allows devices to communicate with one another enabling rich peer–to–peer social applications.

Advantages:

- Minimal-electrical power consumption & vast library resources.
- Handset makers like it because they can use and customize the platform without paying a royality.
- Developers like it because they know the platform has & is not locked into any one vendor that may go under or be acquired.
- Parts of one application can be used in another ways not originally developed by the developer,Even can replace built in components with new versions,which will lead to new creativity in mobile space.
- Programs are isolated fromeach other by multiple layers of security,which provides a level of system stability.
- Open - Android allows you to access core mobile device functionality through standard API calls.
- All applications are equal - Android does not differentiate between the phone's basic and third-party applications -- even the dialer or home screen can be replaced.
- Breaking down boundaries - Combine information from the web with data on the phone -- such as contacts or geographic location -- to create new user experiences.
- Fast and easy development - The SDK contains what you need to build and run Android applications, including a true device emulator and advanced debugging tools.

Limitations: Android operating system, this system also has shortcomings,

- Security -Making source code available to everyone inevitably invites the attention of black hat hackers.
- Open Source - A disadvantage of open-source development is that anyone can scrutinize the source code to find vulnerabilities and write exploits.
- Login-Platform doesn't run on an encrypted file system and has a vulnerable login.
- Incompetence – Google's dependence on hardware and carrier partners puts the final product out of their control.

New Versions:

With all upcoming applications and mobile services Google Android is stepping into the next level of Mobile Internet. Android participates in many of the successful open source projects. That is, architect the solution for participation and the developers will not only come but will play well together. This is notable contrast with Apple and other companies, where such architecture of participation is clearly explained.Android 4.0 makes common actions more visible and lets users navigate with simple, intuitive gestures. Refined animations and feedback throughout the system make interactions engaging and interesting. An entirely new typeface optimized for high-resolution screens improves readability and brings a polished, modern feel to the user interface. Android 4.0 delivers a refined, unified UI for phones and tablets and introduces innovative features for users and developers. Android 4.0 simple, beautiful, and beyond smart.Android may well find itself competing against the forthcoming Nokia touch screen phones and may be even the iPhone 2.

References:

http://developer.android.com/index.html

IT 049

SOCIAL ISSUES IN TECHNOLOGY TRANSFER

J. K. Lahir
Professor, ASM's IBMR, Pune
Maharastra, India
Email:- jklahir@asmedu.org
Mob: +91-9822095283

Nitin P.Ganeshar

Asst. Professor, ASM's IBMR, Pune
Maharastra, India
Email:-nitin.ganeshar@asmedu.org

Mob:-9970231869

Abstract:

The advancements in science and technology have always been associated with the social issues. Whether technology has contributed to the question about a specific problem, or has been its cause, new scientific discoveries and technologies never get away from inspection. Every field of venture must come to provisions with the social implications of these advancements. Social issues related to technology in specific areas Such as Survival, Computers, Internet, Work, Artificial intelligence, Environment. Every field of endeavor must come to terms with the social implication of these advances and innovations. No researching topic of social significance can afford to ignore the technology or science.

The advancement in technology is mostly used for the destruction purposes and effective means of detecting and policing breaches, maybe it can not, as such the survival becomes the important issue. Due to technological development many issues are arising, such as issues related to social and human welfare, urban and regional challenges, education and employment. The issues related to socio-demographic and technology-transfer is the problems associated to an aging population, and also the implementation of new socially oriented technologies.

Keywords:

Internet – Network of network.

Artificial Intelligence – To replace the human activities and enhancing human skills.

Environment - surroundings of an object.

Venture - An undertaking of chance, risking of something upon an event which can not be foreseen with certainty.

Socio-demographic -Relating to, or involving a combination of social and demographic factors.

INTRODUCTION

The social environment is governed by the conditions including customs, morals, values and demographic characteristics of the society in which the technology transfer has taken place. Dealing with different expectations of people is the crux of cross cultural management, this has to be effective in this area .The organization will need to develop a broader understanding of the social environment when technology transfer is required. The business ethics demands a close analysis of these prerequisites as also the social and business structures that are conducive to moral principles.

Right from beginning the advances and innovations in technology, and sciences are interconnected with the social issues of that age. Every field of endeavor must come to terms with the social implication of these advances and innovations .No researching topic of social significance can afford to ignore the technology or science. These effects are consequential more acutely then ever. The social issues are more environmentally based but that is indicative of contemporary world .The issues such as nuclear or plant hazardous wastes and examining human gene therapy, the search for extra terrestrial intelligence etc. It uses following part of approaches:

- Brief background on the technology and science
- Historical contact showing social, ethical, economical are other issues evolved.
- Various points of view on each issue.
- The business and society have a symbiotic relationship, with business acting the role of the junior partner. It is strongly affected by the surrounding social, political, ethical, moral and ecological environment.

Society exerts influence on business in several ways:

- The society legal system
- Political & government regulations

- Public Opinion and Community attitudes
- Influence on the nature of business growth & development.

The concept of morality and ethics held in that particular society determines its development. Technology consists of technique and knowledge.

Technology is considered to be the body of scientific and engineering knowledge which can be applied in the design of products and process or in the search of new knowledge.

The word technology is derived from two Greek words: 'Techne' and 'logos'. Techne means technique, the skill to make something. 'logos' means discussion or knowledge of some entity

CHARASTERISTIC OF TECHNOLOGY:-

A) Resource:

1) Resources are: Money, Time, and People.

2) Technology Development (TD) consumes resources.

3) Resource may include collateral assets.

B) Transferability:

Technology and knowledge Transfer is not easy. Knowledge is always scarce and sticky. Between the sender and receiver during the communication, there is no perfect correspondence.

C) Opportunity:

Technology development takes place when human beings perceive an opportunity for improvement due to intrinsic or economic reasons.

D) Appropriability:

In a certain instances, TD is for economic motives, wherein individuals will pursue development only to the extend that there is a reasonable assurance; the fruits of their labor will flow back to be developers.

Schnepp *et al.* define TT as "…a process by which expertise or knowledge related to some aspect of technology is passed from one user to another for the purpose of economic gain". On the other hand, the Intergovernmental Panel on Climate Change (IPCC) provides a broad definition of TT, in terms of a set of processes, "covering the flows of know-how, experience and equipment, for mitigating and adapting to climate change amongst different stakeholders such as governments, private sector entities, financial

institutions, Non-Governmental Organizations (NGOs) and research/education institutions".

SOME OF TH MOST OFTEN-CITED CONSIDERATIONS ARE LISTED BELOW:

1) Technology transfer includes both "soft" and "hard" elements of technology.
2) Technology transfer should address both climate change adaptation and mitigation technologies.
3) The role of governments is crucial, even though the transfer of technology usually involves many Stakeholders, and is a complex process.

4) Technology transfer should enhance the endogenous capacity of developing and transition countries to develop and implement change response technologies.

PROCESS OF TECHNOLOGY TRANSFER:-

1) Technology management:
Technology management focuses in competitive advantages for the organization who have effective means to fulfill the needs and the requirements in the market.

Requirement need find effective ←------ **Technology** -----→ focus upon competitive advantages

Of customer means to fulfill **management** creating for organization

2) Technology change

The successful technology transfer takes place in the following way we have existing problem and new problem we solve this by some old solution and new solution together we apply to solve the new problems by this method technology transfer process . The technology transfer as focused in international cooperation agreements between developed and developing states. Such agreements may be related to:

*Infrastructure

*Agriculture development

*International cooperation in the fields of:

Research.

Education

Employment

Transport

FORMS OF TECHNOLOGY TRANSFER (TT)

There are two forms of TT: Internalized and Externalize.

Internalized the technology transfer by MNC's is mostly internalized as such it does

Externalize technology transfer needs lot of efforts and capital.

Technology Transfer Levels

1. Innovative level

It is the level which is characterized by innovative skills and R&D.

2. Operational Level

It involves basic manufacturing skills.

3. Adaptive level

Adaptive level is the self reliance level in this the imported technologies are adapted and improved and design skills and complexity of engineering are learned.

4. Duplicative Level

It refers to the investment capability required to expand capacity and to integrate foreign technologies.

Technology Flow Channels

Foreign Investment ← Technology Flow channels → Joint Production Agreement

Important channels for the flow of technology are foreign investment, technology license agreement and joint venture

TECHNOLOGY TRANSFER METHOD

There are many methods for technology transfer it depends on the variety of organization depending on the following:

A) Type of technology
B) Nature of technology assistance
C) Extend of technology assistance

Following methods are normally adopted for technology transfer:

1) Training of local people

2) Deputation of Technical Expert.

3) Collaboration agreement

4) Supply of machinery and equipments

5) Total plant engineering service contract

6) Turnkey contracts.

7) Joint working agreement

Many a times combination of two or more above methods are under taken however the Turnkey Contracts are the most comprehensive of such combination.

ISSUES IN TRANSFER TECHNOLOGY

There are four important issues in technology transfer

1) Cost
2) Appropriateness
3) Dependence
4) Obsolescence

1) Cost:

To obtain a foreign technology from other countries involves high price in number of cases of FDI (foreign Direct Investment) associated with technology transfer, the net out flow to captain by way of dividend, interest, technical fees and royalty etc has been found much higher than the corresponding flow.

2) Appropriateness:

The important factor to be consider in technology transfer (TT) is physical, economic and social conditions of the developing country in many cases it has been found that TT has been irrelevant or inappropriate to the recipient countries socio-economic priorities and conditions. In many cases the state-of-the-art technology is not transferred to the recipient country.

It is preferable that the collaborators should offer appropriate technology by scaling down their technology at lesser cost to reduce the burden on recipient country most if the collaborators would think of one time benefit rather than long term benefit with the recipient country.

3) Dependence:

It is imperative that the recipient country of technology transfer is heavily relining on foreign technology which leads to technological dependence on that country.

4) Obsolescence:

It is normally observed that there is a tendency to transfer outdated technology to the developing countries therefore the recipient country would not have the advantage of latest technology and would still technologically lag behind and normally they are exploited. Many countries salvage technology that is obsolescent in the advance country even-though they posses more advance technology.

The technology absorption in Indian company has been rather slow the main causes are:

1) Lack of incentives to improve

2) Resistance to change

3) Lake of infrastructural facilities

4) Lake of research and development (R&D) facilities.

5) Poor ability to absorb technology.

SOCIAL ISSUES

The advancements in science and technology have always been associated with the social issues. Whether technology has contributed to the question about a specific problem, or has been its cause, new scientific discoveries and technologies never get away from inspection. Every field of venture must come to provisions with the social implications of these advancements. Similarly, anyone researching on a topic of social significance cannot overlook the science or technology, if present, that plays a part in it. In our modern culture, these effects are felt more intensely than ever.

Following are the social issues related to technology in specific areas-

- **Survival:** The advancement in technology is mostly used for the destruction purposes as the technology does not have a persistent ethical code to which its people almost commonly subscribe, and an effective means of detecting and policing breaches, maybe it cannot as such the survival becomes the important issue.
- **Computers:** The 'Computer' has become an integral part of the society. It is imperative that computer has to be an issue for technology management. Many are still in the state of confusion whether they are good or bad. It depends on the way we use it.
- **Internet:** Internet is universally accessible technology. It is being used along with the computer. As we all know that computer and internet has become one of the basic requirements for us. Internet can be used for both constructive and destructive purposes. This becomes a social issue for the business & the government.
- **Work:** All of us should think about the answer to the question 'how does a less specialized society offer everyone of us with job? The problem solvers will always have work. But only the type of the problems they solve will alter. For example, robotics guarantees that there will be few unskilled or repetitive jobs in the fourth civilization. But they will not take over complete work until and unless we design them to do so. All these concerns provide a professionalizing of work and not its elimination.
- **Artificial intelligence:** This is an interesting field regarding the use of technology, which is being used to replace various human activities, enhancing human skills etc. The purpose behind this is that it should be for noble cause.
- **Environment:** Environment is very important for human society. The technology has to be environment friendly. Technology helped the humankind to discover the nuclear power which both a boon and bane to us The products and wastes while generating nuclear power is harmful for the environment.

Due to technological development many issues are arising, issues related to social and human welfare, urban and regional challenges, and education and employment. The issues related to socio-demographic and technology-transfer, are:

· Problems associated to an aging population, and also the implementation of new socially oriented technologies.

· Surveys indicating varying demographic arrangement and the position of technical education.

· Effect of new technologies on human behaviors.

A social issue in Science & Technology is a variable resource and it provides a good overview and valuable background information about a variety of topics.

SUMMERY

Right from beginning the advances and innovations in technology, and sciences are interconnected with the social issues of that age. Every field of endeavor must come to terms with the social implication of these advances and innovations no researching topic of social significance can afford to ignore the technology or science. The advancements in science and technology have always been associated with the social issues. Social issues related to technology in specific are as Survival, Computers, internet, work, artificial intelligence, environment. The important Ethical issues are as Plagiarism, Piracy, Ergonomics/health issues, Digital divide, Gender, Nanotechnology. The moral issues confronted in technology management are environmental protection, consumer safety, and a question of loyalty. An environmental impact analysis is the analysis of possible positive or negative impact that a proposed activity may have on the environment. It also consists of the social and the economy aspects. The purpose of analysis is to ensure that decision maker consider the ensuing environmental impact when deciding whether to processed with a project.EIA involved a technical evaluation that would lead to objective decision making. There are various methods available to carry out EIAs some are industry specific and some generic methods such as industrial products, genetically modified plant, Fuzzy arithmetic.EIA can benefit threatened species conservation . Instead of concentrating on the direct effects of a propose projects on its local environment the benefits of EIA, can be used as landscape approach which focuses on much broader relationships between the entire population of a variety of species.

REFERENCES

1. Technology Management by C. S. V. Murthy Himalaya publications.
2. Business Ethics (concepts and practices) by Dr. Agalgatti and Krishna Nirali Prakashan.
3. Production and Operations Managements-Chase Aquilqno Jacobs (Tata McGraw Hill).
4. Managing engineering and technology (An introduction to management for engineers) (Daniel L Baboock.) prentice hall.
5. Management of technology (The key to competitiveness and wealth creation); Jarlk M Khalil (Tata McGraw Hill).
6. Management information system CSV Murthy Himalaya, Publishing house.
7. Sites:-http://techtransfer.energy.gov

IT 050

Virtual Keyboard

Prof.Jyothi Salunkhe

INTRODUCTION

Virtual Keyboard is just another example of today's computer trend of '**smaller and faster**'.Many alternate forms like speech recognition, hand writing recognition devices had come. The new virtual keyboard technology uses sensor technology and artificial intelligence to let users work on any surface as if it were a keyboard.

The software and hardware part recognizes the typed characters and pass it to the computer. Virtual Keyboard, being a small, handy, well-designed and easy to use application, turns into a perfect solution for cross platform multilingual text input.

QWERTY KEYBOARDS

QWERTY is the most common keyboard layout on English-language computer and typewriter keyboards.

INSIDE THE KEYBOARD

The processor in a keyboard has to understand several things that are important to the utility of the keyboard, such as:

- Position of the key in the **key matrix**.
- The amount of **bounce** and how to filter it.
- The speed at which to transmit the **typematics**.

ciThe microprocessor and controller rcuitry of a keyboard

A Look at the Key Matrix

VIRTUAL KEYBOARD

Virtual Keyboard is just another example of today's computer trend of "smaller and faster".Virtual Keyboard uses sensor technology and artificial intelligence to let users work on any surface as if it were a keyboard. Virtual Devices have developed a flashlight-size gadget that projects an image of a keyboard on any surface and let's people input data by typing on the image.

The device detects movement when fingers are pressed down. Those movements are measured and the device accurately determines the intended keystrokes and translates them into text.The Virtual Keyboard uses light to project a full-sized computer keyboard onto almost any surface, and disappears when not in use.Used with Smart Phones and PDAs, it provides a practical way to do email, word processing and spreadsheet tasks, allowing the user to leave the laptop computer at home.

VIRTUAL KEYBOARD TECHNOLOGY

This system comprises of three modules:

1. The sensor module

2. IR-light source and

3. The pattern projector

Sensor module:

The Sensor Module serves as the eyes of the Keyboard Perception technology .The Sensor Module operates by locating the user's fingers in 3-D space and tracking the intended keystrokes, or mouse movements

I**R-light source**:

The Infrared Light Source emits a beam of

infrared light . This light beam is designed to overlap the area on which the keyboard pattern projector or printed image resides. This helps in recognizing the hand movements and the pressing of keys .

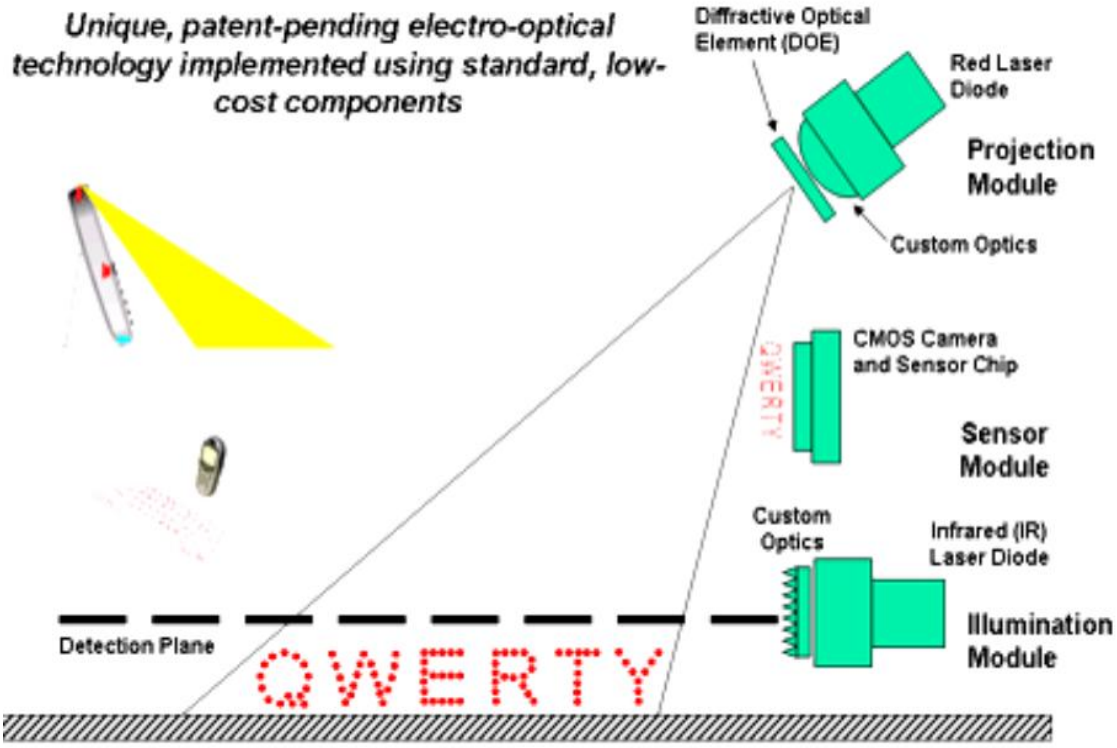
The pattern projector:

The Pattern Projector or optional printed image presents the image of the keyboard or mouse zone of the system. This image can be projected on any flat surface. The projected image is that of a standard QWERTY keyboard, with all the keys and control functions as in the keyboard.

VIRTUAL DEVICES

For some input devices there is a complimentary form where they can also be displays. This is called input-output duality. Input/output duality is accomplished by designing two separate technologies into one integrated package.

Recently a new class of device has started to emerge which is conceptually rooted in exploiting this input/output duality. They can be called Projection/Vision systems, and/or Projection/Scanning or Projection/Camera technologies.



Technical Terms

Projection Module – A template is created of the desired interface and is projected onto the adjacent interface surface. This template is produced by illuminating a specially designed highly efficient holographic optical element with a red diode laser.

Micro-illumination Module – An infa-red plane of light is generated just above, and parallel to the interface surface. This light is invisible to the user and hovers a few millimeters above the surface.

Sensor Module – Reflected light from user interactions with the interface surface is passed through an infa-red filter and imaged onto a CMOS image sensor in the sensor module.

DIFFERENT VIRTUAL KEYBOARDS

There are different types of virtual keyboards, manufactured by various companies which provide different levels of functionalities . The different types of virtual keyboards are:

1)Developer VKB 2) Canesta 3) Sense board Technologies 4) KITTY 5) InFocus

Developer VKB:

Siemens Procurement Logistics Services ,Rechargeable batteries similar to those in cell phones. As a Class 1 laser, the output power is below the level at which eye injury can occur.

Canesta :

In the Canesta Keyboard the same laser is also used to scan the projection field and extract 3D data. They also have a chip set, *Electronic Perception Technology,* which they supply for 3rd parties to develop products using the projection/scanning technology.

Sense board Technologies

Here the image is not projected on to any surface. The Senseboard SB 04 technology is an extreme case of a hybrid approach. The sensing transducer is neither a laser scanner nor a camera. Rather, it is a bracelet-like transducer that is worn on the hands which captures hand and finger motion.

Sensors made of a combination of rubber and plastic are attached to the user's palms in such a way that they do not interfere with finger motions.

KITTY

KITTY, an acronym for Keyboard-Independent Touch-Typing, is a Finger mounted keyboard that uses touch typing as a method of data entry.

InFocus

They do not use laser technology. This has that advantage of delivering high quality color images with a mature technology like video camera. It has the disadvantage of larger size, lower contrast, and higher power requirements, compared to laser projection systems.

Advantages

- It can be projected on any surface or you can type in the plain air.
- It can be usefull in places like operation theaters where low noise is essential.
- The typing does not require a lot of force. So easing the strain on wrists and digits.
- The Virtual Keyboard is not restricted to the QWERTY touch-typing paradigm, adjustments can be done to the software to fit other touch-typing paradigms as well.
- No driver software necessary, It can be used as a plug and play device.
- High battery life. The standard coin-sized lithium battery lasts about eight months before needing to be replaced.

Drawbacks

- Virtual keyboard is hard to get used to. Since it involves typing in thin air, it requires a little practice. Only people who are good at typing can use a virtual keyboard efficiently.
- It is very costly ranging from 150 to 200 dollars.
- The room in which the projected keyboard is used should not be very bright so that the keyboard is properly visible.

Conclusion

- Projection key boards or virtual key boards claim to provide the convenience of compactness with the advantages of a full-blown QWERTY keyboard. The company's Virtual Keyboard is designed for anyone who's become frustrated with trying to put information into a handheld but doesn't want to carry a notebook computer around.
- Canesta appears to be the most advanced in this class of technology and the only one who is shipping product. Other products are KITTY, a finger-mounted keyboard for data entry into PDA's, Pocket PC's and Wearable Computers and KITTY, a finger-mounted keyboard for data entry into PDA's, Pocket PC's and Wearable Computers.

- Thus virtual keyboards will make typing easier, faster, and almost a pleasure.

Reference

- **1. http://www.newscom.com/cgi-bin/prnh**
- **2. www.canesta.com**
- **3. www.procams.org**
- **4. www.billbuxton.com/3state.html**
- **5. www.smarttech.com**
- **6.www.3m.com/us/office/meeting/product_catalog/wd.jhtml**

IT 051

Practitioner Methods for Testing Software Applications designed for Cloud Computing Paradigm

Ms. Vaishali Jawale[1]
Assistant Professor
ASM's Institute of Computer Studies Pimpri - Pune,
University of Pune, India
vj.istqb@gmail.com[1]

Prof.Asheesh Dixit[2]
Director
ASM's Institute of Computer Studies Pimpri – Pune
University of Pune,India
asheeshdixit@yahoo.com[2]

Abstract:

Cloud computing is basically an Internet-based network made up of large numbers of servers. Clouds contain vast amounts of information and provide a variety of services to large numbers of people. Economically, the main appeal of cloud computing is that customers only use what they need, and only pay for what they actually use. Resources are available to be accessed from the cloud at any time, and from any location via the internet. Because of this, cloud computing has also been called utility computing, or 'IT on demand'. While many companies are approaching cloud computing with cautious optimism, testing appears to be one area where they are willing to be more adventurous. Cloud-based testing introduces a new set of challenges, such as performance related issues, data security and a lack of standards, especially in the public cloud model.

This research paper gives an introduction of cloud testing and focuses on various approaches to test the cloud applications deployed and developed on cloud. This paper also focuses on the comparison of testing the application outside the cloud and testing the same application inside the cloud.

Keywords: Cloud Computing, Public Cloud, Cloud Testing, Utility Computing, Application Performance, Data Security, Deployment.

1.0 Introduction:

Cloud computing has gained a significant amount of attention in the last few years. It includes virtualized hardware and software resources that are hosted remotely and made available on-demand pay-as-you-go using a services model (e.g., SOA). Instead of running or storing applications locally, one can host their application in the cloud and access it from anywhere using a thin client application such as a Web browser. Cloud computing promises to reduce cost by cutting down the need for buying large amount of hardware and software resources. It also promises efficiency, flexibility, and scalability**.[1]**

Cloud computing delivers infrastructure, platform, and software that are made available as subscription-based services in a pay-as-you-go model to consumers. These services are referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) in industries**.[2]**

Service Models

- **SaaS (Software as a Service):** The consumer uses an application, but does not control the operating system, hardware or network infrastructure on which it's running.
- **PaaS (Platform as a Service):** The consumer uses a hosting environment for their applications. The consumer controls the applications that run in the environment (and possibly has some control over the hosting environment), but does not control the operating system, hardware or network infrastructure on which they are running. The platform is typically an application framework.
- **IaaS (Infrastructure as a Service):** The consumer uses "fundamental computing resources" such as processing power, storage, networking components or middleware. The consumer can control the operating system, storage, deployed applications and possibly networking components such as firewalls and load balancers, but not the cloud infrastructure beneath them.

There are three types of cloud deployment models available; however there are is one another type of cloud deployment model known as community cloud which is being used in some instances.

- **Public Cloud**: In simple terms, public cloud services are characterized as being available to clients from a third party service provider via the Internet.The term “public” does not always mean free, even though it can be free or fairly inexpensive to use. A public cloud does not mean that a user’s data is publically visible; public cloud vendors typically provide an access control mechanism for their users. Public clouds provide an elastic, cost effective means to deploy solutions.

- **Private Cloud**: A private cloud offers many of the benefits of a public cloud computing environment, such as being elastic and service based. The difference between a private cloud and a public cloud is that in a private cloud-based service, data and processes are managed within the organization without the restrictions of network bandwidth, security exposures and legal requirements that using public cloud services might entail. In addition, private cloud services offer the provider and the user greater control of the cloud infrastructure, improving security and resiliency because user access and the networks used are restricted and designated.

- **Community Cloud**: A community cloud is controlled and used by a group of organizations that have shared interests, such as specific security requirements or a common mission. The members of the community share access to the data and applications in the cloud.
- **Hybrid Cloud**: A hybrid cloud is a combination of a public and private cloud that interoperates. In this model users typically outsource non-businesscritical information and processing to the public cloud, while keeping business-critical services and data in their control.

So much is said and done about Cloud computing, it is the single largest focal point of the computing infrastructure exist today. Many of the Web Applications are moving to Cloud and perhaps the most important reason for leveraging cloud capabilities is to quickly gain access to hundreds or thousands of computers for computing capacity when needed, performance becomes an essential and integral part, IT organizations are aiming to gain agility in their applications and infrastructure. A number of small to medium-sized IT organizations have migrated to cloud solutions. As a result, cloud testing has become necessary to validate functional system and business requirements. In addition to cloud experience, cloud testing engineers require the knowledge of different types of testing and tools.

Application performance, security and customer loyalty go hand-in-hand. According to 2010-11 World Quality report**[3],** a quarter of all survey respondents indicate that they encountered application performance issues within the first few months of moving to the cloud infrastructure, followed by 20% who experienced security and vulnerability problems.

In this paper, three different Approaches for testing Application on Cloud are suggested,

1. **Section 3.0 Cloud Testing- TYPE A**

[Testing the cloud application which is already deployed on the cloud.] (SaaS)(Black Box Testing)

2. **Section 4.0 Cloud Testing-TYPE B**

[Testing the cloud application during development on the cloud.](SaaS,PaaS,Iaas)(White Box, Black Box)

3. **Section 5.0 Cloud Testing-TYPE C**

[Testing the cloud application during development off the cloud. Development and testing well going on iteratively on premises as well as on cloud until the application is ready to use.](Saas)(White Box, Black Box)

The rest of this paper is organized as follows: first, a general description about Cloud computing, cloud testing, cloud services and cloud types are presented. This section ends with a brief overview of 3 different approaches for testing cloud application. Following that, in background section comprehensive details related to different conventional types of testing as well as types of testing have to be considered for applications that are developed for working on cloud environment are described. Section 6 presents the overall comparison of all three approaches of cloud application testing. Finally, the paper ends with brief conclusive remarks and discussion on future research directions.

2.0 Background:

Software Testing is the process of executing a program or system with the intent of finding errors. Or, it involves any activity aimed at evaluating an attribute or capability of a program or system and determining that it meets its required results. Software testing, depending on the testing method employed, can be implemented at any time in the development process. However, most of the test effort occurs after the requirements have been defined and the coding process has been completed. As such, the methodology of the test is governed by the software development methodology adopted.

2.1 Testing Non Cloud Application:

Traditionally we are testing the non cloud application by using different conventional types of testing.**[4]**

Black box testing – Internal system design is not considered in this type of testing. Tests are based on requirements and functionality.

White box 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.

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 harnesses

Integration testing – Testing of integrated modules to verify combined functionality after integration. Modules are typically code modules, individual applications, client and server applications on a network, etc. This type of testing is especially relevant to client/server and distributed systems.

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.

System testing – Entire system is tested as per the requirements. Black-box type testing that is based on overall requirements specifications, covers all combined parts of a system.

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.

Sanity testing - Testing to determine if a new software version is performing well enough to accept it for a major testing effort. If application is crashing for initial use then system is not stable enough for further testing and build or application is assigned to fix.

Regression testing – Testing the application as a whole for the modification in any module or functionality. Difficult to cover all the system in regression testing so typically automation tools are used for these testing types.

Acceptance testing -Normally this type of testing is done to verify if system meets the customer specified requirements. User or customer does this testing to determine whether to accept application.

Load testing – It's a performance testing to check system behavior under load. Testing an application under heavy loads, such as testing of a web site under a range of loads to determine at what point the system's response time degrades or fails.

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.

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.

Usability testing – User-friendliness check. Application flow is tested, Can new user understand the application easily, Proper help documented whenever user stuck at any point. Basically system navigation is checked in this testing.

Install/uninstall testing - Tested for full, partial, or upgrade install/uninstall processes on different operating systems under different hardware, software environment.

Recovery testing – Testing how well a system recovers from crashes, hardware failures, or other catastrophic problems.

Security testing – Can system be penetrated by any hacking way. Testing how well the system protects against unauthorized internal or external access. Checked if system, database is safe from external attacks.

Compatibility testing – Testing how well software performs in a particular hardware/software/operating system/network environment and different combination s of above.

Alpha testing – In house virtual user environment can be created for this type of testing. Testing is done at the end of development. Still minor design changes may be made as a result of such testing.

Beta testing – Testing typically done by end-users or others. Final testing before releasing application for commercial purpose.

2.2 Testing Cloud Application:

In addition to the types described in section 2.1 the following types of testing have to be considered for applications that are developed for working on cloud environment**.[5]**

Application changes and on-premise interfaces - Moving an application to the cloud may require changes to the application to suit with the environment

available on the cloud platform. For example, if the application is running on Solaris on-premise and the operating systems available on the cloud platform are RedHat Linux and Suse Linux, the application will require additional testing on the cloud platform. In addition, requirements such as authentication against on-premise active directory and new interfaces built with on-premise systems lead to additional areas for application testing.

Data migration - The data migrated from on-premise to cloud to take advantage of the storage services available on the cloud platform for additional storage needs, backup and archival requires new test cases to be developed as part of test planning.

Security - Organizations may decide to enforce access to application features utilizing on-premise user directories for authentication and authorization. In addition, web applications being migrated from with-in the firewall to a public internet on cloud require transport security mechanisms. Data stored in cloud storage may have to be encrypted for security and compliance needs. Security testing of these new features has to be incorporated in to overall test planning.

Performance - The application needs to be load and stress tested on the cloud platform to ensure that system response is as per SLAs. As cloud platforms are often used for consolidating infrastructure in multiple regions to a specific region, special attention is needed to ensure that system performance is at optimum levels. This will require testing the latency in system response in comparison to response from on-premise environment for taking appropriate steps for performance improvement. At the time of designing the performance test cases, following important factors are given adequate consideration:

1. **Capacity Handled:** refers to the maximum amount, which can be handled, or can be produced, or can completely occupy the entity.

2. **Reliability:** Ensure that it consistently produces the desired results for a given set of functions.

3. **Stress:** Should be able to move the system or its component beyond the specified limits of performance expectations.

4. **Response time:** Should be able to discover the total time taken by a response to be received after initiating a request for it.

5. **Bandwidth:** Provide consideration to the bandwidth, or a particular quantum of data passing across a physical entity.

6. **Efficiency:** Provide consideration to the efficiency in terms of the ratio of quantum of data processed to the amount of resources utilized to process.

7. **Recovery:** Refers to the ability of the system to return to normal processing after removal or reduction of the load. It includes estimation of time period for such a recovery

8. **Availability:** Provide the constant (or near constant) availability of the application and data. The client should test the availability for an extended duration e.g. 24 hours or 48 hours. We can do this by simulating a virtual user performing the identified business transactions in repeated iterations for the entire duration. After the test finishes, the client should analyze the test log to determine the following errors:

a. Any errors related to missing or unavailable resources e.g. server unavailable or page not found
b. Request time outs

9**. Scalability** - Applications are usually migrated to cloud to take advantage of the elasticity features provided by cloud platforms for quick ramp up/down of computing resources to align with actual demand. The ability of the cloud environment to meet the on-demand needs of the application requires careful planning of the test scenarios and load for performance testing.

10**. Availability and Disaster Recovery** - Load balancers and elastic computing features for auto-provisioning play a critical role in ensuring high availability on the cloud. The ability of the cloud environment to withstand peak load and server failures has to be tested prior to releasing the application for producing use. Cloud platforms such as Amazon EC2 allow placing multiple instances in different availability zones in a region to protect applications from failure of a single location .The availability of the application has to be tested by simulating the failure of an application instance in a location. Also, data recovery mechanisms in place have to be tested for ensuring proper disaster recovery.

Following are the some basic steps in cloud testing, **[6]**

Steps In Cloud Testing

Here we are proposing 3 different methods of testing the cloud application. While testing we considered some scenarios for that application. In every scenario the testing will be done step by step. We are testing the application for functional as well as non functional behavior. But we are more focusing on non functional behavior of the application in cloud environment.

3.0 Cloud Testing- TYPE A

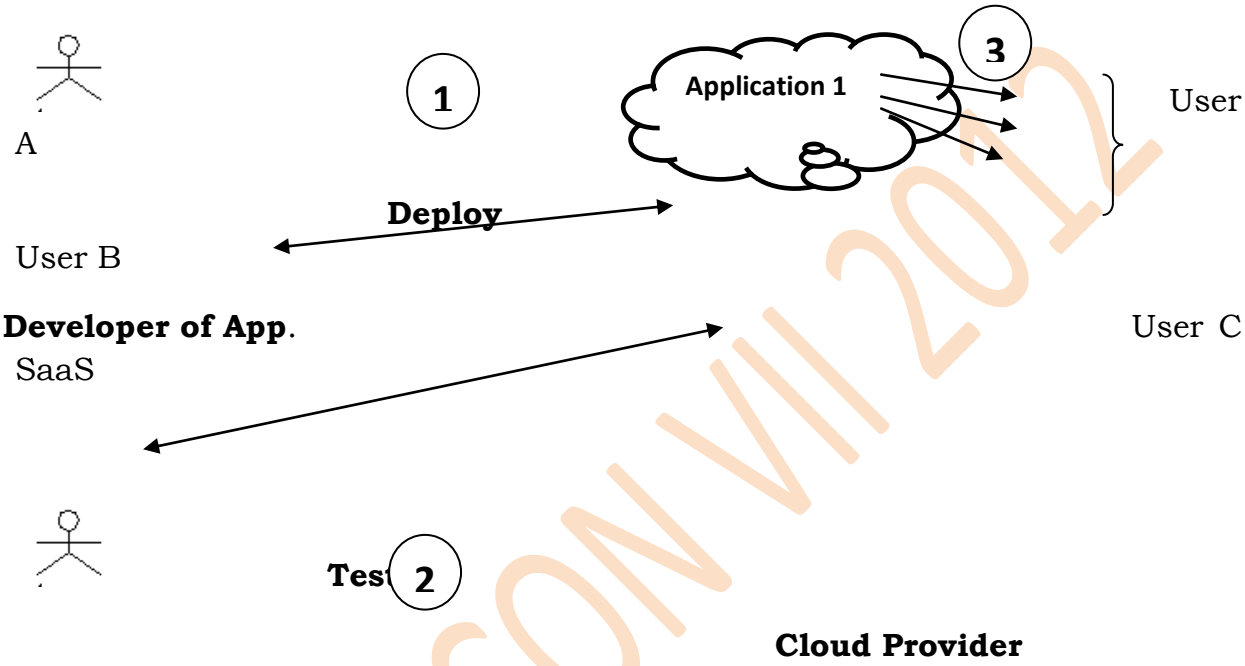


Fig.1 Cloud Testing TYPE A

In this approach we are testing the application which is already deployed on cloud. Let us consider above scenario as shown in fig 1.In this case the application was developed on premises by developer and then it is deployed on the cloud by cloud provider. Multiple users can use this application by using SaaS.After deployment tester will starts testing the application. Here there will not be any changes to the functional test planning and execution between a 'On Premise' and 'Cloud Application', however the non functional test planning will differ and hence needs to be addressed**[7].** So the testing will be focused on various non functional testing types described in section 2.1 as well as 2.2. Tester will design test cases and run them on application1 for single user as well as multiple users. If he finds any bug, he will prepare test report and send it to the developer. Developer will fix the bugs and again redeploy that application on cloud. This is an iterative process. It repeats

until the deployed application will successfully used by users without any error.

4.0 Cloud Testing- TYPE B

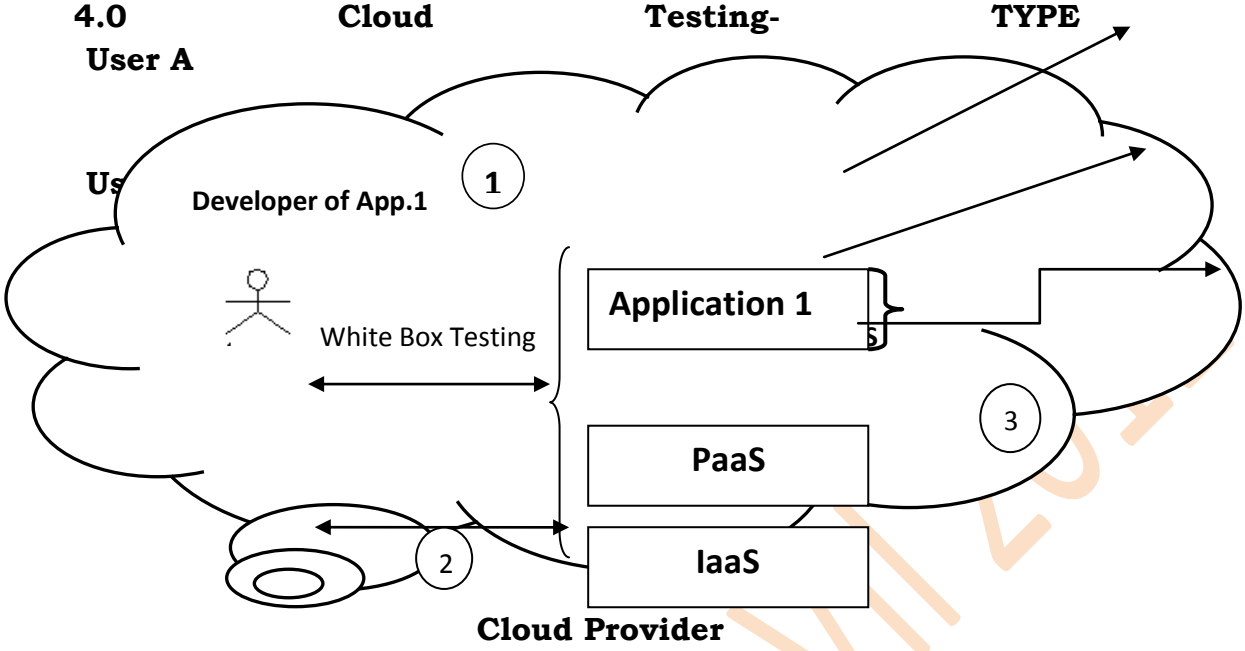


Fig.2 Cloud Testing: TYPE B

In this approach we are testing the application during development on cloud. During testing cloud based applications we keep in mind that the application is completely built on the cloud itself. Let us consider above scenario as shown in fig 2.In this case the application was developed on cloud by developer by using PaaS,IaaS and then it is tested on the cloud by tester. Multiple users can use this application by using Saas. During development developer will starts white box testing of the application. This testing focuses on procedural details i.e., internal logic of a program. Using white box testing, the developer can design the test cases as follows**,[8]**

a) Guarantee that all independent paths within a module have been exercised at least once.
b) Exercise all logical decisions on their true and false sides.
c) Execute all loops at their boundaries and within their operational bounds.
d) And exercise internal data structures to ensure their validity

After completion of all functional tests by developer, tester comes in a role for non functional testing. So he will test it by using various non functional testing types described in section 2.1 as well as 2.2. He will design test cases and run them on application1 for single user as well as multiple users. If he finds any bug, he will prepare test report and send it to the developer. Developer will fix the bugs and again send it to the tester. This is an iterative process. The total process will executes on cloud. It repeats until the cloud

application will successfully used by users without any error in cloud environment.

5.0 Cloud Testing- TYPE C

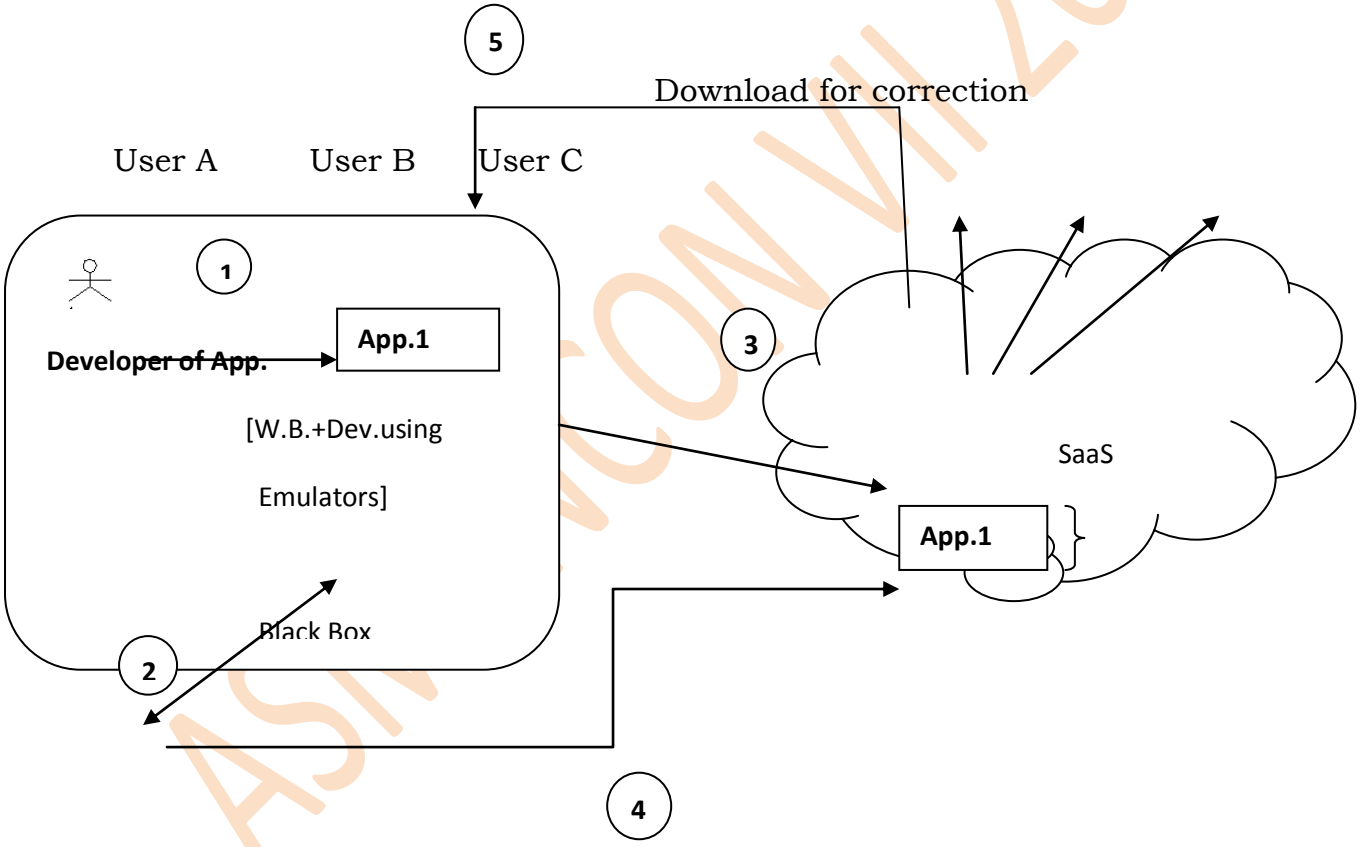


Fig.3 Cloud Testing: TYPE C

Some of the traditional and emerging Cloud-based application services include social networking, web hosting, content delivery, and real time instrumented data processing. Each of these application types has different composition, configuration, and deployment requirements. Quantifying the

performance of provisioning (scheduling and allocation) policies in a real Cloud computing environment for different application models under transient conditions is extremely challenging because:

(i) Clouds exhibit varying demands, supply patterns, system sizes, and resources (hardware, software, network);

(ii) Users have heterogeneous, dynamic, and competing QoS requirements;

(iii)Applications have varying performance, workload and dynamic application scaling requirements.

The use of real infrastructures on cloud for benchmarking the application performance (throughput, cost benefits) under variable conditions (availability, workload patterns) is often constrained by the rigidity of the infrastructure. Hence, this makes the testing extremely difficult. Further, it is tedious and time-consuming to re-configure benchmarking parameters across a massive scale Cloud computing infrastructure over multiple test runs. Such limitations are caused by the conditions prevailing in the Cloud-based environments that are not in the control of developers and tester of application services. Thus, it is not possible to perform benchmarking experiments in repeatable, dependable, and scalable environments using real-world Cloud environments**.[2]**

A more viable alternative is the use of Emulators. Emulator-based approaches offer significant benefits to IT companies (or anyone who wants to offer his application services through clouds) by allowing them to:

(i) test their services in repeatable and controllable environment;

(ii) Tune the system bottlenecks before deploying on real clouds;

(iii) Experiment with different workload mix and resource performance scenarios on simulated infrastructures for developing and testing adaptive application provisioning techniques.

In this approach we are testing the application during development on developer premises. Let us consider above scenario as shown in fig 3.In this case the application was developed on premises by developer and then testing is done by tester using emulators on their local computer. After this emulator testing that application will deploy on the cloud and ready to use for single as well as multiple users.

Tester has to configure the emulators for testing the application. An emulator is a software program that aims to replicate the functions of a specific piece of hardware or software. Emulator tests the cloud-based

applications in a number of different environments before it deploys the application to the live production environment. To ensure high quality of cloud applications under development, developer have to test a unit without waiting for other units to be available, being able to detect and remove software faults at a lower cost comparing to do so at a later stage**[9].** To conduct white box testing on cloud applications, a practical way is to employ various desktop based cloud environment emulators and Storage Emulators, which enable developers to run and test their cloud applications locally rather than testing them after deployment. In addition, developers also need to provide various test inputs. Most test cases (which are written for testing a unit that interacts with the cloud environment) begin with a manual step of preparing environment setup and these test cases must run against a local cloud environment simulator**.[10]**

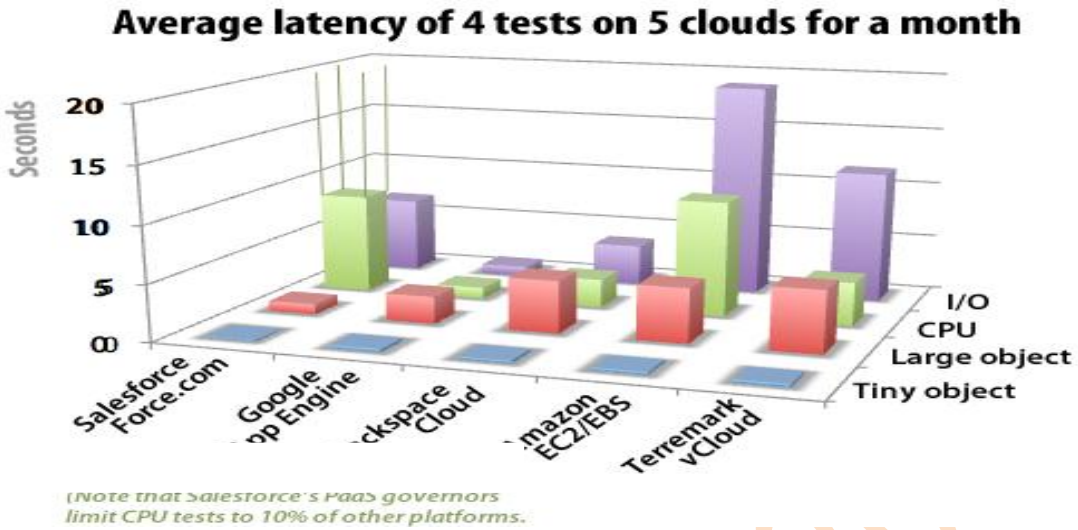
6.0 Comparison of Testing Types:

In this section we compare above mentioned three different approaches to test the cloud application which is develop outside the cloud as well as inside the cloud. These three approaches are, testing the application outside the cloud and testing the application inside the cloud. Compared to maintaining an in-house test environment cloud-based testing offers lower costs, more flexibility enhanced collaboration and several other benefits. The Following table shows the key points of our comparison.

Sr No.	**Parameter**	**Cloud Testing: TYPE 1**	**Cloud Testing: TYPE 2**	**Cloud Testing: TYPE 3**
1	**Use of Hardware & software**	Hardware and Software Procurement	No need of installation and set up	Proper configuration of emulators has to be done.
2	**Costs**	High Maintenance Costs	Reduce testing costs	High Emulators cost
3	**Use of Simulation**	Inability to simulate multiple geographies	No need of Simulation	Creation of real word situations through

				simulation of geographically distributed load patterns
4	**Test quality**	Normal	Improved	Improved
5	**Time**	It consumes time because all installations, configurations have to be done.	No need to wait to buy and configure servers	It consumes time because emulators has to be configures according to real environment.
6	**Scalability**	Scalability cannot be easily calculated**.**	Cloud testers can quickly and easily scale from thousands to millions of concurrent users	Through Emulators tester can scale millions of users.
7	**Latency Period**	Low	High	High
8	**Support for complex apps**	Expensive and Ineffective	Cheap & Highly effective	Comparatively less Expensive

The Cloud environment forces to test non functional parameters such as availability, security, usability, scalability and performance by equating their importance with functionality testing. According to survey, the performance of the applications to five different clouds is as follows which are monitored for a month. It is discovered that performance varies widely by test type and cloud**.[11]**



7.0 Conclusion:

In this Paper, we explained different approaches to test the application which is going to deploy on cloud, the application which is developed on cloud. We also did the comparison between an in-house test and cloud-based test of an application .By comparing them we concluded some points which shows that, testing the application inside the cloud is having some advantages over testing the application outside the cloud.

While working on these various approaches of testing we notice that, Cloud-based software applications have some additional characteristics compared to non-cloud-based ones. These pose additional challenges but with a systematic, comprehensive approach to test planning, these could be handled.

We need to remember that there is no single or ideal approach for cloud testing. This is primarily due to the fact that when an organization embarks onto cloud testing, various factors like the cloud architecture design, non-functional and compliance requirements, etc., need to be taken into account to ensure successful and complete testing.

8.0 References:

[1] Scott Tilley and Tauhida Parveen, *Florida Institute of Technology*. Software Testing in the Cloud Perspectives on an Emerging Discipline book.

[2] Rodrigo N. Calheiros1, 3, Rajiv Ranjan2, Anton Beloglazov1, César A. F. De Rose3, and Rajkumar Buyya1. CloudSim: A Toolkit for Modeling and

Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms.

[3] Neha Thakur,ISTQB certified professional, QA Lead. Performance Testing In cloud: A Pragmatic Approach.A white paper in STC 2010.

[4] http://www.softwaretestinghelp.com/types-of-software-testing

[5] Rajagopal Sattaluri. Testing Considerations for Application Migration to Cloud Computing, an article for cloud expo in Cloud Computing Journal.

[6] http://en.wikipedia.org/wiki/Cloud_testing

[7] Srinivasan Sundara Rajan. Non-Functional Testing for Cloud, Testing for Cloud QoS,An article in Cloud Computing Journal.

[8] http://www.itswtech.org/Lec/Software Engineering/ 2013.pdf

[9] http://msdn.microsoft.com/en-us/library/ff728592.aspx. Moving Applications to the Cloud, 2nd Edition,msdn library.

[10] Linghao Zhang, Tao Xie, Nikolai Tillmann,Peli de Halleux, Xiaoxing Ma, Jian Lv. Environment Modeling for Automated Testing of Cloud Applications. In Proc of IEEE **ISSN:** 0740-7459

[11] A Bitcurrent study on the performance of cloud computing platforms .June, 2010 Cloud computing performance.

http://www.webmetrics.com/landingpage/bitcurrentcloud2/The_Performance_of_Clouds_Summary.pdf

IT 052

"RESEARCH AND ANALYSIS ON NEW APPROACH TO WEB APPLICATIONS AND ARCHITECTURAL STYLE FOR AJAX"

SHAILESH TEJRAM GAHANE

ASSISTANT PROFESSOR

ASM'S INSTITUTE OF COMPUTER STUDIES PIMPRI - PUNE

UNIVERSITY OF PUNE, INDIA

E-MAIL: shaileshmca2007@gmail.com

ABSTRACT

A new breed of web application, AJAX, is emerging in response to a limited degree of interactivity in large grain stateless web interactions. At the heart of this new approach lies a single page interaction model that facilitates rich interactivity. We have studied and experimented with several AJAX frameworks trying to understand their architectural properties.

In this paper, we summarize three of these frameworks and examine their properties and introduce the architectural style of AJAX. We describe the guiding software engineering principles and the constraints chosen to induce the desired properties. The style emphasizes user interface component development, and intermediary delta communication between client/server components, to improve user interactivity and ease of development. In addition, we use the concepts and principles to discuss various open issues in AJAX frameworks and application development.

KEYWORDS

AJAX, Framework, Web Interaction, Software Engineering.

1. INTRODUCTION

Over the course of the past decade, the move from desktop applications towards web applications has gained much attention and acceptance. Within

this movement, however, a great deal of user instructiveness has been lost. Classical web applications are based on a multi page interface model [4], in which interactions are based on a page-sequence paradigm. While simple and elegant in design for exchanging documents, this model has many limitations for developing modern web applications with user friendly human computer interaction.

Recently there has been a shift in the direction of web development. A new breed of web application, AJAX (Asynchronous JavaScript and XML) [6], is emerging in response to the limited degree of interactivity in large grain stateless web interactions. At the heart of this new approach lies a single page interface model that facilitates rich interactivity. In this model, changes are made to individual user interface components contained in a web page, as opposed to refreshing the entire page. Thanks to the momentum of AJAX, single page interfaces have attracted a strong interest in the web application development community [7]. After the name AJAX numerous frameworks and libraries have appeared, many web applications have adopted one or more of the ideas underpinning AJAX [8], and an overwhelming number of articles in developer sites and professional magazines have appeared.

Adopting AJAX based techniques is a serious option not only for newly developed applications, but also for existing web sites if their user friendliness is inadequate. A software engineer considering adopting AJAX [8], however, is faced with a number of challenges.

- What are the fundamental architectural differences between designing a legacy web application and an AJAX web application?
- What are the different characteristics of AJAX frameworks?
- What do these frameworks hide?
- Is there enough support for designing such applications?
- What problems can one expect during the development phase?
- Will there be some sort of convergence between the many different technologies?
- Which architectural elements will remain, and which ones will be replaced by more elegant solutions?

Addressing these questions calls for a more abstract perspective on AJAX web applications. Despite all the attention the technology is receiving in the web community, there is a lack of a coherent and precisely described set of architectural formalisms for AJAX enabled web applications. In this paper we explore whether concepts and principles as developed in the software architecture research community can be of help to answer such questions [9]. In particular, we propose different architectural style for AJAX applications, and study to what extent this style can help in addressing our questions [7].

Web interaction designers can't help but feel a little envious of our colleagues who create desktop software. Desktop applications have a richness and responsiveness that has seemed out of reach on the Web. The same simplicity that enabled the webs rapid proliferation also creates a gap between the experiences we can provide and the experiences users can get from a desktop application. Web interaction designers can't help but feel a little envious of our colleagues who create desktop software. Desktop applications have a richness and responsiveness that has seemed out of reach on the web. The same simplicity that enabled the Web's rapid proliferation also creates a gap between the experiences we can provide and the experiences users can get from a desktop application.

That gap is closing. Take a look at Google Suggest. Watch the way the suggested terms update as you type, almost instantly. Now look at Google Maps, Zoom in. Use your cursor to grab the map and scroll around a bit. Again, everything happens almost instantly, with no waiting for pages to reload. Google Suggest and Google Maps are two examples of a new approach to web applications that we at Adaptive Path have been calling AJAX. The name is shorthand for Asynchronous JavaScript + XML [7], and it represents a fundamental shift in what's possible on the Web.

2. AJAX FRAMEWORKS

Web application developers have struggled constantly with the limits of the HTML page sequence experience, and the complexities of client-side JavaScript programming to add some degree of dynamism to the user interface. Issues regarding cross-browser compatibility are, for instance, known to everyone who has built a real-world web application. The rich user interface (UI) experience AJAX promises comes at the price of facing all such problems. Developers are required to have advanced skills in a variety of web technologies, if they are to build robust AJAX applications. Also, much effort has to be spent on testing these applications before going in production. This is where frameworks come to the rescue. At least many of them claim to, because of the momentum AJAX has gained, a vast number of frameworks are being developed. The importance of bringing order to this competitive chaotic world becomes evident when we learn that almost one new framework per day is being added to the list of known frameworks. We have studied and experimented with several AJAX frameworks trying to understand their architectural properties. [8] We summarize two of these frameworks in this

section. Our selection includes a widely used open source framework called "Echo2" and the web framework offered by Google called GWT [7]. These two frameworks are major players in the AJAX market, and their underlying technologies differ substantially.

2.1 ECHO2 FRAMEWORK

Echo2 is an open-source AJAX framework which allows the developer to create web applications using an object oriented, UI component-based, and event-driven paradigm for Web development. Its Java Application Framework provides the APIs (UI components, property objects, and event/ listeners) to represent and manage the state of an application and its user interface. All functionality for rendering a component or communicating with the client browser is specifically assembled in a separate module called the "Web Rendering Engine" [8]. The engine consists of a server-side portion (written in Java/J2EE) and a client-side portion (JavaScript). The client/server interaction protocol is hidden behind this module and as such, it is entirely decoupled from other modules. Echo2 has an Update Manager which is responsible for tracking updates to the user interface component model, and for processing input received from the rendering agent and communicating it to the components. The Echo2 Client Engine runs in the client browser and provides a remote user interface to the server-side application. Its main activity is to synchronize client/server state when user operations occur on the interface. A ClientMessage in XML format is used to transfer the client state changes to the server by explicitly stating the nature of the change and the corresponding component ID the change has taken place on. The server processes the ClientMessage, updating the component model to reflect the user's actions. Events are fired on interested listeners, possibly resulting in further changes to the server-side state of the application. The server responds by rendering a ServerMessage which is again an XML message containing directives to perform partial updates to the DOM representation on the client [8].

2.2 GWT (GOOGLE WEB TOOLKIT) FRAMEWORK

Google has a novel approach to implementing its AJAX framework, Google Web Framework (GWT). Just like Echo2, GWT facilitates the development of UIs in a fashion similar to AWT or Swing and comes with a library of widgets that can be used. The unique character of GWT lies in the way it renders the client-side UI. Instead of keeping the UI components on the server and communicating the state changes, GWT compiles all the Java UI components to JavaScript code (compile-time). Within the components the developer is allowed to use a subset of Java 1.4 API to implement needed functionality. GWT uses a small generic client engine and, using the compiler, all the UI

functionality becomes available to the user on the client. This approach decreases round-trips to the server drastically. The server is only consulted if raw data is needed to populate the client-side UI components. This is carried out by making server calls to defined services [9]. The services (which are not the same as Web Services) are implemented in Java and data is passed both ways over the network using serialization techniques.

3. DEFINING AJAX FEATURES

AJAX is not a technology. It is really several technologies, each flourishing in its own right, coming together in powerful new ways [5].

AJAX incorporates:

- Standards based presentation using XHTML and CSS [1].
- Dynamic display and Interaction using the Document Object Model (DOM) [2].
- Data Interchange and Manipulation using XML and XSLT.
- Asynchronous Data Retrieval using XMLHttpRequest.
- JavaScript binding everything together.

The classic web application model works like this: Most user actions in the interface trigger an HTTP request back to a web server. The server does some processing retrieving data, crunching numbers, talking to various legacy systems and then returns an HTML page to the client. It's a model adapted from the Web's original use as a hypertext medium, but as fans of The Elements of User Experience know, what makes the Web good for hypertext doesn't necessarily make it good for software applications [5].

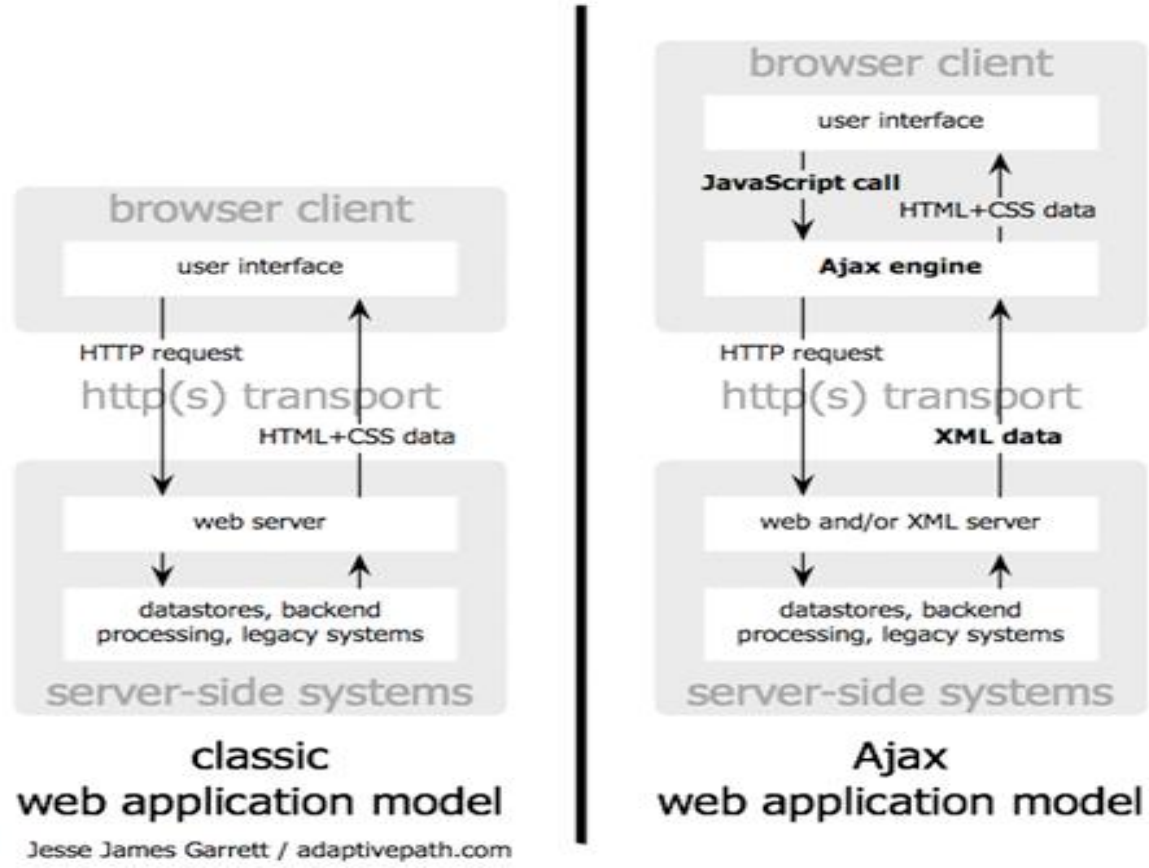


Figure: The traditional model for web applications (left) compared to the AJAX model (right).

4. WHEN WE USE AJAX

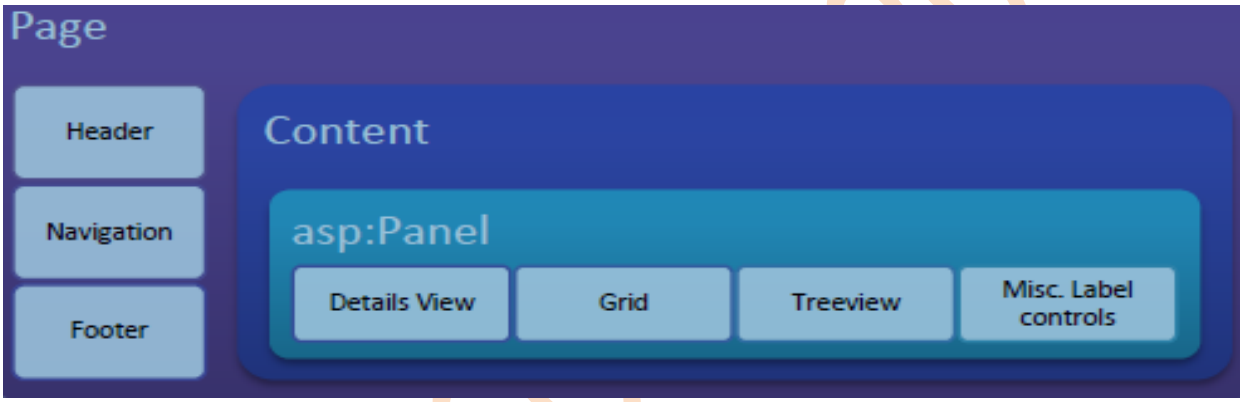
With tools like ASP.NET AJAX [3], it is not hard to find places in your application that can benefit from instant Ajaxification. In general, any user experience, a more holistic approach than user interface in your application that does not involve navigation is a good candidate for AJAX.

Common examples of these interactions are:

- A form that validates values with some server process.
- A drop down list that loads values in response to another element's action.
- Voting or rating input elements.
- Multi-tab interfaces.
- Any grid operations. (Such as sorting, selecting, editing, filtering, etc.)

It is important that you identify actions in your application that can be ajaxified at a very granular level. In other words, if you have a page on which most of the contents are updated after a PostBack, applying AJAX to the page will not provide much (if any) benefit. Remember from our earlier examination of AJAX in part's one and two of this series that AJAX is good at quickly updating small portions of the page. If you apply your ajaxification too broadly, you'll lose many of the underlying benefits.

Let's look at a practical example. Let's say we have page that conceptually looks like this:



If we decide to ajaxify all of the actions on this page that normally use PostBacks, we might be tempted (especially if we're using ASP.NET AJAX UpdatePanels) to simply ajaxify the "asp:Panel". After all, that would automatically convert any action from the Treeview, Grid, or the Details View into an AJAX Callback and allow those actions to update any of the other controls in the Panel. Doing that, though, would probably eliminate most of the value we could extract from AJAX.

5. WHERE DO THE BENEFITS OF AJAX COME FROM?

In business, decision makers are interested mainly in how information technology can reduce costs, or make better use of information assets. The benefits of AJAX seem to come more out of the cost-containment arena than the latter [6].

5.1 POTENTIALLY MEASURABLE BENEFITS

These are benefits that can be measured and expressed in terms of dollars and cents without much difficulty. Regardless of the quality of AJAX UI, you will look to these metrics to estimate value.

They include:

- **Time spent waiting for data to be transmitted:** Time is money. Over many repetitions, the time employees spend waiting for the page to load can add up to significant costs.
- **Time spent completing a particular task:** Increased efficiency in the user interface can often mean that time is saved at the task level, offering opportunities for concrete cost savings there.
- **Bandwidth consumed for the entire task:** The cost of bandwidth does not increase linearly, but does increase as the company invests in larger capacity internet connections and new hardware to accommodate greater server loads. A firm's cost structure for bandwidth depends on the scale of their operation and these capital investment needs. That being said, the cost of bandwidth can be measured if this cost structure is known. If repetitious tasks consume a lot of bandwidth, these costs can escalate dramatically. The amount of bandwidth consumed also has implications for time savings.

5.2 HARD TO QUANTITY BENEFITS

Some of the benefits associated with good user interfaces are qualitative and difficult to measure precisely. This is not to imply they are not of financial value, but many business benefits are hard to quantify and seasoned IT managers will know intuitively they can translate into significant bottom line savings [6]. For AJAX, the opportunities for streamlining the interface are limited only by our imaginations, and it should also be noted that it's still possible to design a terrible UI with AJAX that does not benefit from any of the following:

- **Steps to complete a task:** Reducing the number of steps has implications for the amount of time consumed but also for the number of opportunities for error. Fewer errors mean cost savings down the road when these errors would have to be manually corrected.
- **Familiar user interface:** Quite often these days, Web-based applications are used to replace desktop applications that had superior user interfaces. The benefits of offering users a similar or even just a familiar user interface to

what they use on the desktop means lower training costs, fewer errors, and greater initial productivity.

- **Improved application responsiveness:** More responsive applications can improve productivity not just by reducing "wait," but by promoting a more fluid, uninterrupted workflow. In a responsive application, users can move rapidly from one action to another as quickly as they can visualize the workflow. Less responsive applications can defeat the user's workflow visualization by forcing them to continually wait for program information.

6. HOW AJAX IS DIFFERENT

An AJAX application eliminates the start-stop-start-stop nature of interaction on the web by introducing an intermediary, “An AJAX Engine” between the user and the server. It seems like adding a layer to the application would make it less responsive, but the opposite is true [8].

Instead of loading a webpage, at the start of the session, the browser loads an AJAX engine written in JavaScript and usually tucked away in a hidden frame. This engine is responsible for both rendering the interface the user sees and communicating with the server on the users behalf. The AJAX engine allows the users interaction with the application to happen asynchronously independent of communication with the server [3]. So the user is never staring at a blank browser window and an hourglass icon, waiting around for the server to do something.

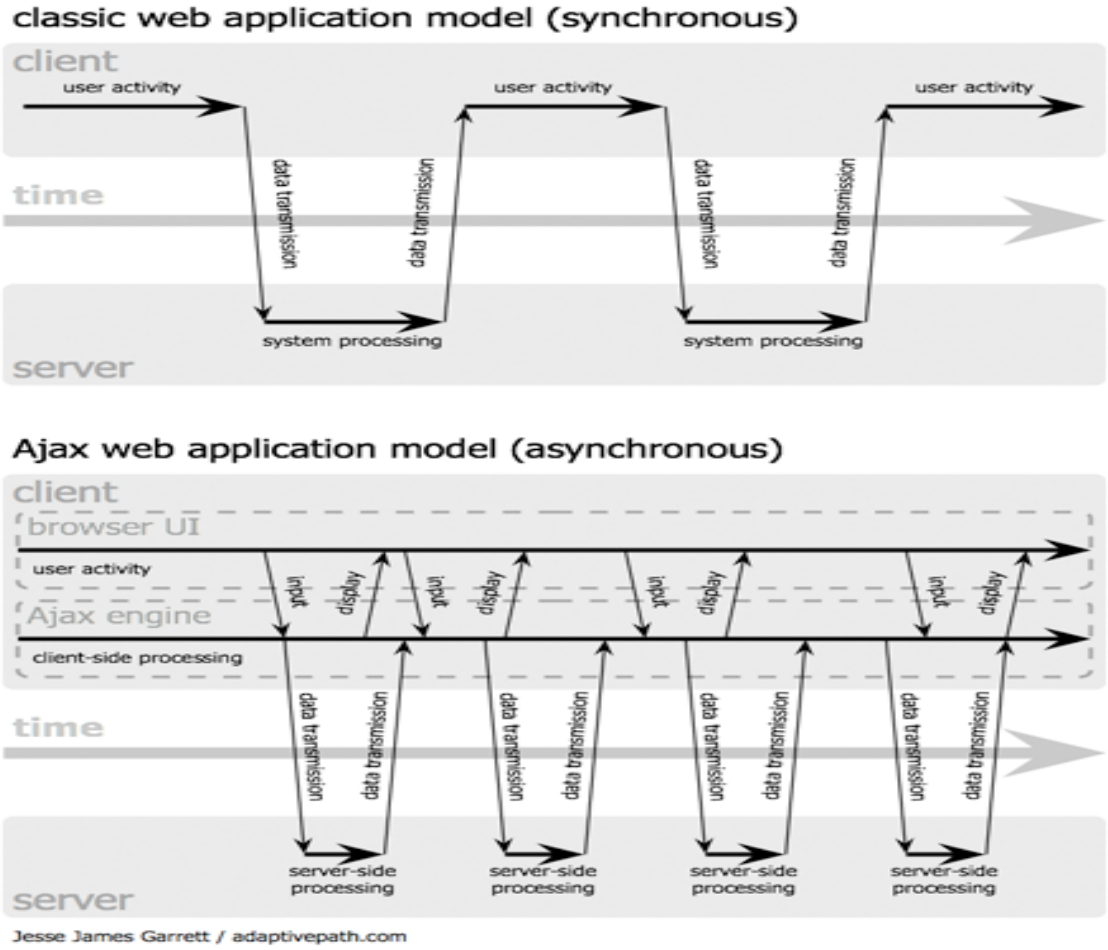


Figure: The Synchronous Interaction Pattern of a Traditional Web Application (Top) Compared With The Asynchronous Pattern of An AJAX Application (Bottom).

Every user action that normally would generate an HTTP request takes the form of a JavaScript call to the AJAX engine instead. Any response to a user action that doesn't require a trip back to the server such as simple data validation, editing data in memory and even some navigation the engine handles on its own. If the engine needs something from the server in order to respond if it's submitting data for processing, loading additional interface code, or retrieving new data, the engine makes those requests asynchronously, usually using XML without stalling a user's interaction with the application [3].

7. TRADITIONAL MODEL VS AJAX MODEL FOR WEB APPLICATIONS

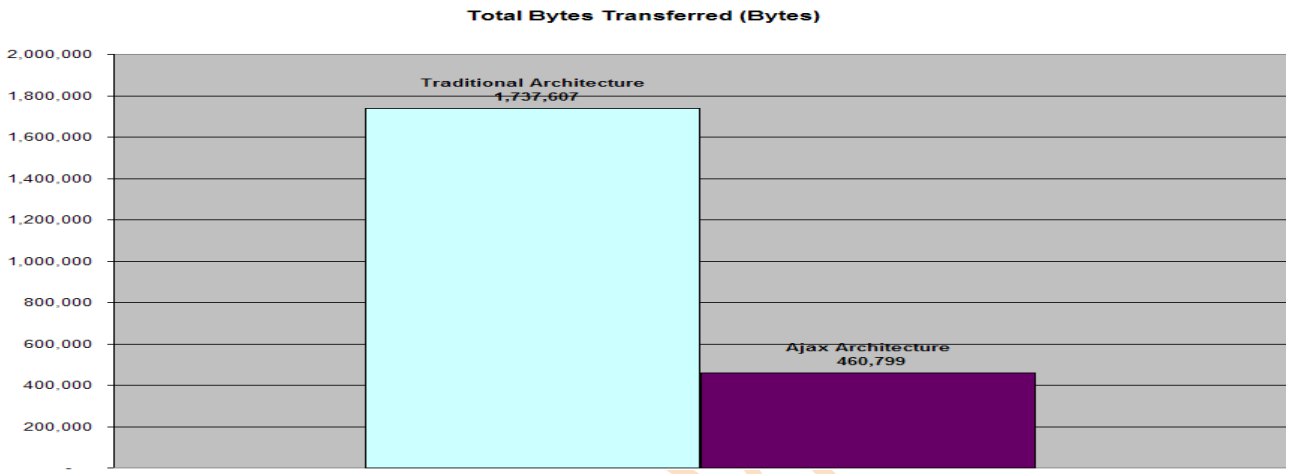
The metrics taken for this study were bytes transferred, total time consumed by the task, and Microsoft Fiddler's estimation of the time it would take to transfer the same data in the same number of HTTP requests to a location on the US West Coast from our position [5].

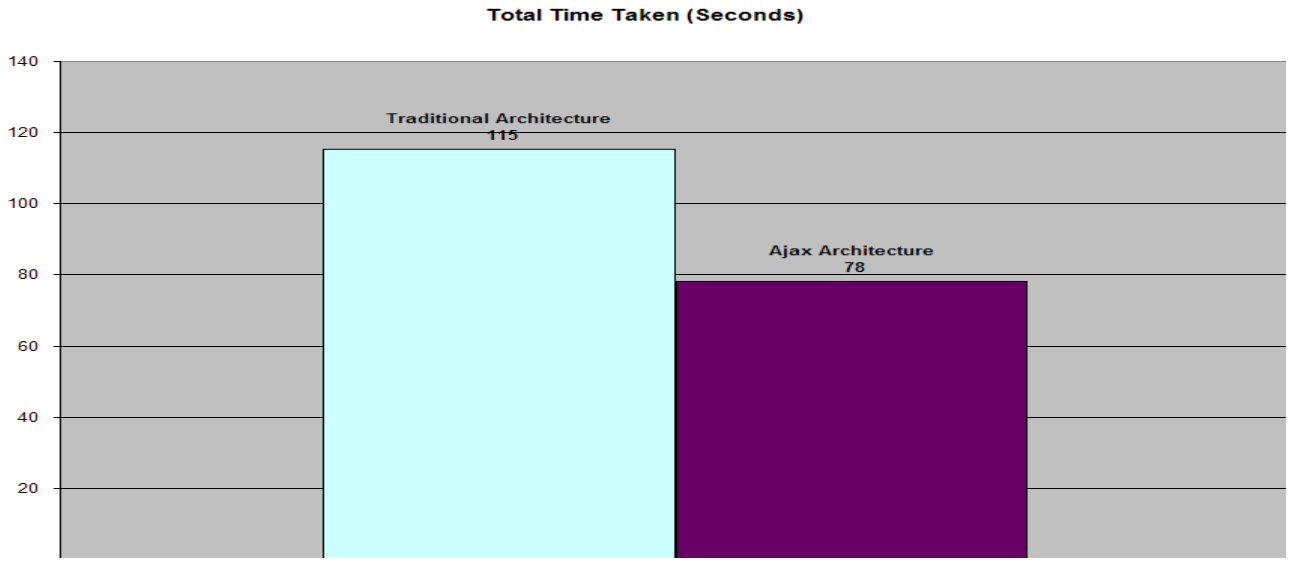
	Traditional Model (Average)	**AJAX Model (Average)**	**Performance Increase**	**Performance Increase (%)**
Bytes Transferred:	1,737,607	460,799	1,276,809	73%
Time (seconds):	115	78	36	32%
Estimated Transmission Time to US West Coast (56k) (Seconds)	293.45	94.44	199.01	68%

As expected, in every test there were significant performance improvements across the board in the AJAX version. The greatest improvements were in bandwidth and network traffic efficiency gains. This resulted in improved responsiveness, allowing the user to move more quickly through the application. Overall, there was a 73% improvement in the number of bytes transferred in the AJAX application over the traditional version. This was due to the fact that server requests were made only for the data that was needed, not for the entire page. Microsoft Fiddler also predicted that there would be a 68% overall time improvement in transferring that information to a location on the US West Coast given the number of HTTP requests, bytes transferred, the latency of that connection, and the bandwidth afforded by a simple 56k modem.

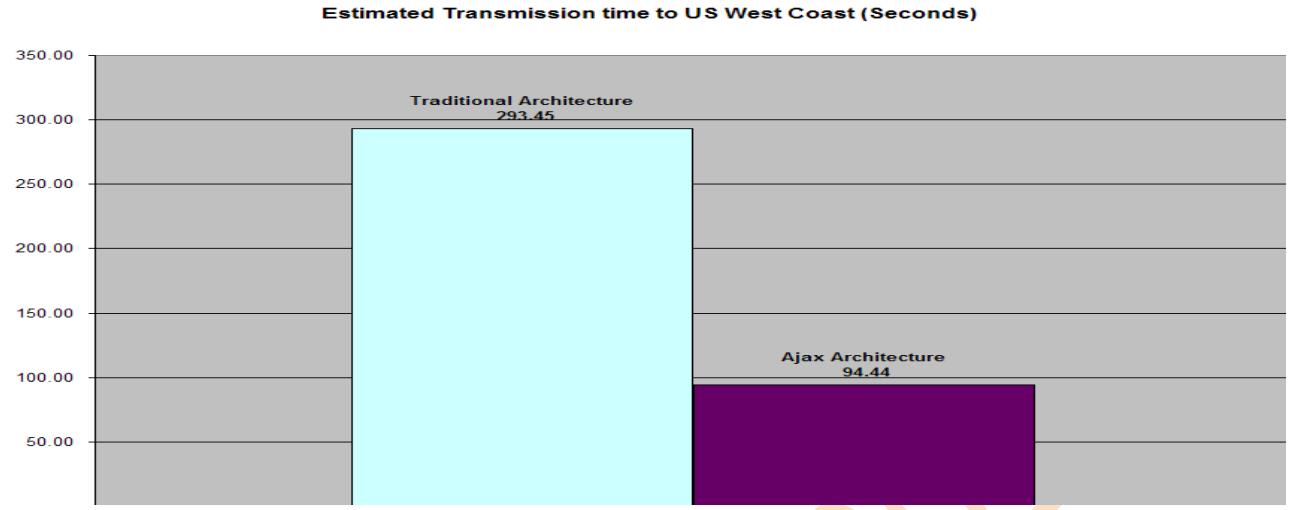
More importantly than bandwidth considerations was the direct benefit to users who saved on average 32% of the time required to complete the tasks in the AJAX version. Had those users been connecting to a remote location

instead of a local server, Fiddler predicted the time savings could have been much more dramatic (around 68%). We can take these numbers now and extrapolate to see how large enterprises could be affected when these actions are repeated over and over [7].





1



8. CONCLUSION

In this paper we have discussed, an architectural style for AJAX. The contributions of this paper are in two research fields: web application development and software architecture. From a software architecture perspective, our contribution consists of the use of concepts and methodologies obtained from software architecture research in the setting of AJAX Internet applications. From a web engineering perspective, our contribution consists of the capturing the guiding software engineering principles that practitioners can use when we constructing and analyzing AJAX frameworks as well as applications. The style is based on an analysis of several of such frameworks, and we have used it to address various design tradeoffs and open issues in AJAX applications.

Although every new technology should be greeted with a healthy amount of skepticism, there are clearly demonstrable, quantifiable advantages to using AJAX architecture in a Web application. These cost savings originate primarily from time savings, but also from reductions in bandwidth requirements. A representative test case showed that a business can save between 500 and 2,800 man hours per year. Although the benefits of improved application architecture extend beyond mere time savings, when included in the decision making process, an ROI approach such as this can help make a solid business case for AJAX.

9. REFERENCES

[1] http://www.asp.net/ajax

[2] http://www.w3schools.com/Ajax/Default.Asp

[3] http://msdn.microsoft.com/en-us/magazine/cc163354.aspx

[4] http://code.google.com/webtoolkit/doc/latest/ReleaseNotes.html

[5] http://ajaxian.com/archives/echo2-new-framework-built-around-ajax

[6] J. Garrett. AJAX: A new approach to web applications. Adaptive path, 2005.

[7] R. Asleson and N. T. Schutta. Foundations of AJAX. Apress, 2005.

[8] D. J. Barrett, L. A. Clarke, P. L. Tarr, and A. E.Wise. A framework for event-based software integration. ACM Trans. Software Engineering Methodol., 5(4):378–421, 1996.

[9] R. Fielding. Architectural styles and the design of network-based software architectures. PhD thesis, UC, Irvine, Information and Computer

IT 053

AN OVERVIEW OF UNSTRUCTURED DATA AND ITS PROCESSING TECHNIQUES

Megha Joshi[1]
Assistant Professor, Institute of Computer Studies,
Pune University, Pune, India
Email: meghhajoshhi@gmail.com
Ph: 7387999531

Vinita Yadav[2]
Assistant Professor, Institute of Computer Studies,
Pune University, Pune, India
Email: vinitay2011@gmail.com
Ph: 9975895813

Abstract: *The unstructured data problem is not easy to describe. Managing such a data is recognized as one of the major unsolved problems. Managing information and associated cost are the major issues for a business while working with unstructured data. Because approach of handling unstructured data is different as that of structural data in terms of their tools and techniques; so it commands more consideration. Understanding data usage and consumption trends again are factors to be considered. In this paper we are trying to present an overview of Unstructured data and some of techniques to handle it. Text Mining in which Patterns are extracted from natural language rather than databases and Multimedia Mining which deals with the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia files. If these multimedia files are analyzed, useful information to users can be revealed.*

Key-Words: - audio mining, image mining, multimedia mining, text mining, unstructured data, video mining.

1. INTRODUCTION

The majority of whole information potential data which we used in organization, enterprises or in day to day life is unstructured. This can include speech, images, e-mail messages, complicated reports, presentations, voice mail, and video. Due to lack of "explicit semantics" which is required for application to interpret the information as required by end-user application or by human it is referred as "Unstructured". Getting valuable information from unstructured data is a complex task and requires usage of special techniques. Handling unstructured data is different than structural data

because of difference in their tools and techniques. Moreover identifying old and relevant data is another issue. Hence right technology is highly desirable to handle this type of data.

Basic Categorization of Unstructured Data

- Bitmap Objects: Essentially non-language based, such as image, video or audio files.
- Textual Objects: Based on a written or printed language, such as Microsoft Word documents, e-mails or Microsoft Excel spreadsheets.

Both of these object types may be seen as data classification, but the technology and methodology for harnessing relevant information from bitmap objects is still in its infancy. Today most of the technologies are available to addresses textual objects [2].

2. HANDLING UNSTRUCTURED DATA

Handling Unstructured data is a Complex task. IBM introduced **Unstructured Information Management Architecture (**UIMA) which is an open, industrial-strength, scalable and extensible **platform** for creating, integrating and deploying unstructured information management solutions from combinations of semantic analysis and search components [5]. It provides a common framework for processing this information to extract meaning and create structured data about the information. The UIMA specification defines platform-independent data representations and interfaces for text and multi-modal analytics. The principal objective of the UIMA specification is to support interoperability among *analytics*.

Several commercial solutions are also available for analyzing and understanding unstructured data for business applications. This includes products from companies like SAS, IxReveal, Inxight and SPSS, as well as more specialized offerings such as Attensity360 and Sysomos, which focuses on analyzing unstructured social media data.

Several ways that can be used to abstract useful information from unstructured data. We here are going to discuss two of such mining techniques in brief.

2.1 TEXT MINING

As we see a huge portion of data is stored in text databases such as archives and digital libraries, news articles, research papers, books, image data banks

used in life sciences, email messages and Web pages. Therefore Text Mining has become essential theme in data mining.

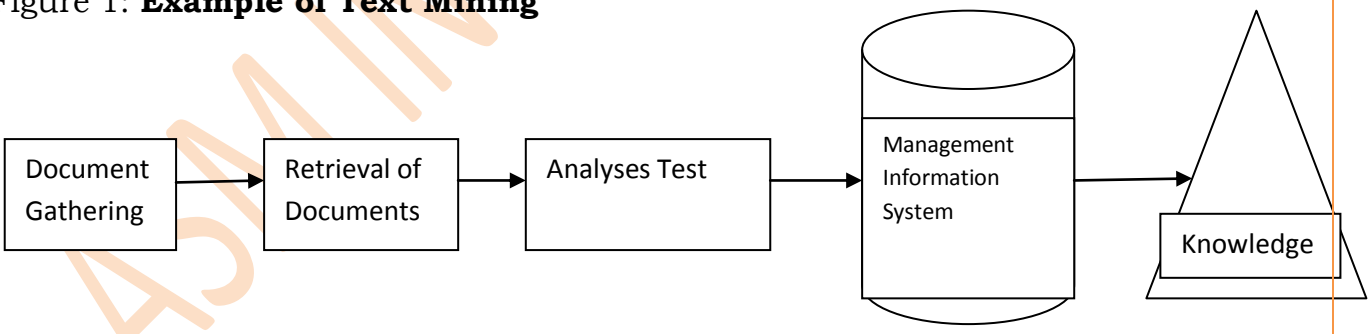
Text-mining, also known as Knowledge discovery from text or Text Data Mining in which information is automatically extracting from usually large amount of unstructured textual resources . Patterns are extracted from natural language rather than databases. The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases.

Although the scope of Text mining are rapidly increasing but the main Text Mining applications are most often used in the sector of Academics, Pharmaceuticals, Government, Insurance and Banking Sector,

There are many approaches to text mining, which can be classified from different perspectives based on inputs in data mining process and tasks to be performed. In general there are three major approaches:

a) The keyword-based- Here input is a set of key words or terms that occurs frequently and then finds association or correlation relationships among them
b) The tagging approach where the input is set of tags
c) The information-extraction approach where the input is a semantic information such as event or facts [3]

Figure 1: **Example of Text Mining**



Source: Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, Journal of emerging technologies in web intelligence Vol. 1 (no. 1, august 2009)

2.1.1 A TEXT MINING SYSTEM IS COMPOSED OF THREE MAJOR COMPONENTS (SEE FIGURE 2):

Information feeders enable the connection between various textual collections and the tagging modules. This component connects to any web site, streamed source (such a news feed), internal document collections and any other types of textual collections.

Intelligent tagging A component responsible for reading the text and distilling (tagging) the relevant information. This component can perform any type of tagging on the documents such as statistical tagging (categorization and term extraction), semantic tagging (information extraction) and structural tagging (extraction from the visual layout of documents).

Business intelligence suite A component responsible for consolidating the information from disparate sources, allowing for simultaneous analysis of the entire information landscape. [1]

Textual data mining and analysis vendors provide analysis tools for unstructured textual objects which are used for clustering documents depending on their contents, Extracting relevant information from a document ,Content wise Classification and organization of documents, Retrieving documents based on the various sorts of information about the document content.

Figure 2: **Architecture of Text-Mining system**

Source: Text mining and information extraction (By Moty Ben-Dov; MDX University, London .Ronen Feldman; Bar-llan University, Ramat-Gun)

2.2 MULTIMEDIA MINING

The other Technique is Multimedia Data Mining or Multimedia Mining which deals with the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia files. Also in retrieval, indexing and classification of multimedia data with efficient information fusion of the different modalities is essential for the system's overall performance [3].

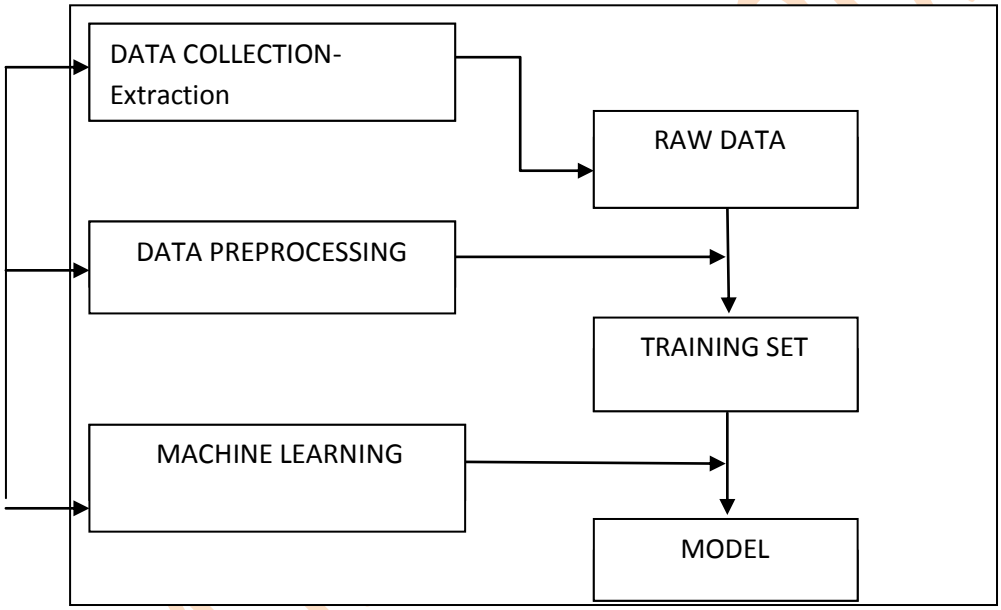
Complexity level is much higher in Multimedia Mining then in Data Mining. The reason being is the large volume of data it contains and subjectivity in understanding the meaning of multimedia contents or due to heterogeneity and variability of the multimedia data.

We present the model of applying multimedia mining in different multimedia types. Data collection is the starting point of a learning system, as the quality of raw data determines the overall achievable performance. Then, the goal of data pre-processing is to discover important features from raw data. Data pre-processing includes data cleaning, normalization, transformation, feature selection, etc. Learning can be straightforward, if informative features can be

identified at pre-processing stage. Detailed procedure depends highly on the nature of raw data and problem's domain. In some cases, prior knowledge can be

Extremely valuable. For many systems, this stage is still primarily conducted by domain experts. The product of data pre-processing is the training set. Given a training set, a learning model has to be chosen to learn from it. It must be mentioned that the steps of multimedia mining are often iterative. [7]

Figure 3: Multimedia Mining Process



Source: S. Kotsiantis, D. Kanellopoulos, P. Pintelas, multimedia mining, My Math Web Portal [portal], Available from: http://www.math.upatras.gr/~esdlab/oldEsdlab/en/members/kotsiantis/multimedia-mining.pdf

The size of the multimedia datasets and its high dimensionality of the feature spaces make the feature extraction a challenging problem. [7]

2.2.1 FEATURES

There are two kinds of features:

a) Description-based – Which uses metadata like caption, keywords, time of creation and size.
b) Content-based – which is based on content of the object

2.2.1.1 FEATURE EXTRACTION FROM TEXT

Text categorization is a conventional classification problem applied to the textual domain. It solves the problem of assigning text content to predefined categories. In the learning stage, the labeled training data are first pre-processed to remove unwanted details and to "normalize" the data. For example, in text documents punctuation symbols and non-alphanumeric characters are usually discarded, because they do not help in classification. Moreover, all characters are usually converted to lower case to simplify matters. The next step is to compute the features that are useful to distinguish one class from another. For a text document, this usually means identifying the keywords that summarize the contents of the document. How are these keywords learned? One way is to look for words that occur frequently in the document. These words tend to be what the document is about. Of course, words that occur too frequently, such as "the", "is", "in", "of" are no help at all, since they are prevalent in every document. These common English words may be removed using a "stop-list" of words during the pre-processing stage. From the remaining words, a good heuristic is to look for words that occur frequently in documents of the same class, but rarely in documents of other classes. In order to cope with documents of different lengths, relative frequency is preferred over absolute frequency. Some authors used phrases, rather than individual words, as indexing terms, but the experimental results found to date have not been uniformly encouraging results. Another problem of text is the variant. Variant refers to the different forms of the same word, e.g. "go", "goes", "went", "gone", "going". This may be solved by stemming, which means replacing all variants of a word by a standard one. [7]

2.2.1.2 FEATURE EXTRACTION FROM IMAGES

Image categorization classifies images into semantic databases that are manually pre-categorized. In the same semantic databases, images may have large variations with dissimilar visual descriptions (e.g. images of persons, images of industries etc.). In addition images from different semantic

databases might share a common background (some flowers and sunset have similar colors). In, the authors distinguish three types of feature vectors for image description:

1) Pixel level features,

2) Region level features, and

3) Tile level features.

Pixel level features store spectral and textural information about each pixel of the image. For example, the fraction of the end members, such as concrete or water, can describe the content of the pixels. Region level features describe groups of pixels. Following the segmentation process, each region is described by its boundary and a number of attributes, which present information about the content of the region in terms of the end members and texture, shape, size, fractal scale etc. Tile level for image features present information about whole images using texture, percentages of end members, fractal scale and others. Moreover, other researchers proposed an information-driven framework that aims to highlight the role of information at various levels of representation. This framework adds one more level of information: the Pattern and Knowledge Level that integrates domain, related alphanumeric data and the semantic relationships discovered from the image data. [7]

2.2.1.3 FEATURE EXTRACTION FROM AUDIO

Audio data play an important role in multimedia applications. Music information has two main branches: symbolic and audio information. Attack, duration, volume, velocity and instrument type of every single note are available information. Therefore, it is possible to easily access statistical measures such as tempo and mean key for each music item. Moreover, it is possible to attach to each item high-level descriptors, such as instrument kind and number. On the other hand, audio information deals with real world signals and any features need to be extracted through signal analysis. The researchers of used only perceptual features such as loudness, brightness, pitch etc. On the other hand, other researchers chose only perceptual features to represent sound clips. Another researcher team used 12 features, as well.

However, some of the most frequently used features for audio classification are:

• Total Energy: The temporal energy of an audio frame is defined by the rms of the audio signal magnitude within each frame.

• Zero Crossing Rate (ZCR): ZCR is also a commonly used temporal feature. ZCR counts the number of times that an audio signal crosses its zero axis.

• Frequency Centroid (FC): It indicates the weighted average of all frequency components of a frame.

• Bandwidth (BW): Bandwidth is the weighted average of the squared differences between each frequency component and its frequency centroid.

• Pitch Period: It is a feature that measures the fundamental frequency of an audio signal. [7]

2.2.1.4 FEATURE EXTRACTION FROM VIDEO

In video mining, there are three types of videos:

a) The produced (e.g. movies, news videos, and dramas),

b) The raw (e.g. surveillance videos etc)

c) The medical video (e.g. ultra sound videos including echocardiogram).

The first stage for mining raw video data is grouping input frames to a set of basic units, which are relevant to the structure of the video. In produced videos, the most widely used basic unit is a shot, which is defined as a collection of frames recorded from a single camera operation. Shot detection methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based. Color or grayscale histograms (such as in image mining) can also be used. To segment video, color histograms, as well as motion and texture features can be used.

Generally, if the difference between the two consecutive frames is larger than a certain threshold

Value, then a shot boundary is considered between two corresponding frames. The difference can be determined by comparing the corresponding pixels of two images. [7]

3. CONCLUSION

This paper describes unstructured data and some of the ways to handle it. Unstructured data is categorized as bitmap and textual objects. We concentrated on Text data and Multimedia. In Text mining if domain specific knowledge is to be extracted it can be seen as a challenge. Because it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts

described in the documents; these methods are computationally expensive. So the focus is to make semantic analysis efficient for large amount of data. And in Multimedia mining if we take case of audio and video mining, issue prevails in combining video and audio information into one comprehensive score. In image mining combining different types of image data is tedious. Further study is already in progress and we are looking after some algorithm to be used for multilingual text refining and to overcome problems of audio/video and image data.

4. REFERENCES:

1. Moty Ben-Dov, Ronen Feldman, Text mining and information extraction, (e-book Chap 38)
Available From: http://vmg.pp.ua/books/%D0%9A%D0%BE%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D1%8B%D0%98%D1%81%D0%B5%D1%82%D0%B8/Mining/text%20mining/%28Mineria%20Textos%29%20Text%20Mining%20and%20Information%20Extraction.pdf

2. Geoffrey Weglarz (2004), Information Management Magazine, Two Worlds of Data – Unstructured and Structured. Issues, September 2004, Available at: http://www.information-management.com/issues/20040901/1009161-1.html , [submitted on Sept 2004]

3. Bojan Ćirić. Working with Unstructured Data, Global data consultancy [www], on 26 November, 2009. Available from: http://www.globaldataconsulting.net/articles/theory/working-unstructured-data [Nov, 2009]

4. Vishal Gupta and Gurpreet S. Lehal, A Survey of Text Mining Techniques and Applications, journal of emerging technologies in web intelligence Vol. 1 (no. 1, august 2009)

5. IBM, Unstructured information management architecture (UIMA), UIMA is an Open, Industrial-Strength Platform for Unstructured Information Analysis and Search. Available from: http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.index.html [Sept 2007]

6. Department of Computer Science & Engineering IIT, Delhi Multimedia mining, Available from: http://www.cse.iitd.ernet.in/~abhinav/datamining/website/ (Feb 2007)

7. S. Kotsiantis, D. Kanellopoulos, P. Pintelas, multimedia mining, My Math Web Portal [portal], Available from: http://www.math.upatras.gr/~esdlab/oldEsdlab/en/members/kotsiantis/multimedia-mining.pdf

8. Cornell University Library, A Survey on Web Multimedia Mining [www] Available from: http://arxiv.org/abs/arXiv:1109.1145 *(Submitted on 6 Se*