Global Journal of Computer Science and Technology

discovering thoughts and inventing future

Volume 10 Issue 5 Version 1.0 Online ISSN: 0975-4172 Print ISSN: 0975-4350

15 Technology Reforming Ideas

highlights

- Optimized Remote Network
- Routing Protocols
- QRS Wave Detection
- Load Balanced Clusters

July 2010

ENG

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 10 Issue 5 (Ver. 1.0)

Global Academy of Research and Development

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://www.globaljournals.org/global-journals-research-portal/guideline/terms-and-conditions/menu-id-260/.

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027

Global Academy of Research and Development

Publisher's correspondence office

Global Journals, Headquarters Corporate Office,
United States

Offset Typesetting

Global Journals, City Center Office,
United States

Packaging & Continental Dispatching

Global Journals, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investers@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)

Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 500 USD (Color)

Editorial Board Members

Chief Author

Dean & Editor-in-Chief (HON.)

Contents of the Volume

From the Chief Author's Desk

We see a drastic momentum everywhere in all fields now a day. Which in turns, say a lot to everyone to excel with all possible way. The need of the hour is to pick the right key at the right time with all extras. Citing the computer versions, any automobile models, infrastructures, etc. It is not the result of any preplanning but the implementations of planning.

With these, we are constantly seeking to establish more formal links with researchers, scientists, engineers, specialists, technical experts, etc., associations, or other entities, particularly those who are active in the field of research, articles, research paper, etc. by inviting them to become affiliated with the Global Journals.

This Global Journal is like a banyan tree whose branches are many and each branch acts like a strong root itself.

Intentions are very clear to do best in all possible way with all care.

Dr. R. K. Dixit
Chief Author
chiefauthor@globaljournals.org

Load Balanced Clusters for Efficient Mobile Computing

GJCST Computing Classification
C.1.3, C.2.4

Dr. P.K.Suri[1] Kavita Taneja[2]

***Abstract*-Mobile computing is distributed computing that involves components with dynamic position during computation. It bestows a new paradigm of mobile ad hoc networks (MANET) for organizing and implementing computation on the fly. MANET is characterized by the flexibility to be deployed and functional in "on-demand" situations, combined with the capability to ship a wide spectrum of applications and buoyancy to dynamically repair around broken links. The underlying issue is routing in such dynamic topology. Numerous studies have shown the difficulty for a routing protocol to scale to large MANET. For this, such network relies on a combination of storing some information about the position of the Mobile Unit (MU) at selected sites and on forming some form of clustering. But the centralized Clusterhead (CH) can become a bottleneck and possibly lead to lower throughput for MANET. We propose a mechanism in which communication outside the cluster is distributed through separate CHs. We prove that the overall averaged throughput increases by using distinct CHs for each neighboring cluster. Although increase in throughput, reduces after one level of traffic rates due to overhead induced by "many" CHs.**

I. Mobile Computing: Vision and Challenges

Mobility originates from a desire to move toward the resource or to move away from scarcity and in rare cases it may be just a nomadic move. Wireless mobile computing faces additional constraints induced by wireless communications and the demand for anytime anywhere communication towards the vision of ubiquitous or pervasive computing. It is accepted that the new parameters in mobile computing [1] are mobility of elements, the limited resources of the Mobile Units (MUs) and the limited wireless bandwidth. The ―mobility” and ―position” has a more significant effect on the development of middleware, simulators and services for the MU than the other parameters. These characteristics can be viewed in a hierarchical fashion where the basic elements influence higher more complicated systems. The mobile computing challenges on the one hand irrevocably handicapped the existing infrastructure in effectively supporting the exponentially rising demands and on the other hand open new avenues and opportunities for Mobile Ad Hoc Network (MANETs). In general, such solutions rely on a combination of storing some information about the position of the MU at selected sites and on forming some form of clustering. The MUs are grouped in distinct or overlapping clusters for the purpose of routing and within the cluster MUs be in touch directly. However, MUs communicate outside the cluster through a centralized MU that is called Clusterhead (CH). CH elected to be part of the backbone for the MANET system and is assigned for communication with all other clusters [2, 3, 4]. This provides a hierarchical MANET system which assists in making the routing scalable. CHs are elected according to several techniques. The CH allows for minimizing routing details overhead from other MU within the cluster. Overlapping clusters might have MUs that are common among them which are called gateways [5]. MANET requires efficient routing algorithm in order to reduce the amount of signaling introduced due to maintaining valid routes, and therefore enhance the overall performance of the MANET system [6,7]. As the CH is the central MU of routing for packets destined outside the cluster in the distinct clustering configuration, the CH computing machine pays a penalty of unfair resource utilization such as battery, CPU, and memory [8]. Several studies [9, 10, 11] have proposed a CH election in order to distribute the load among multiple hosts in the cluster. Our approach extends the same concept of load balancing among CHs too. Section 2 discusses the related work and outlines major challenges while clustering in MANETs, section 3 discusses the multi-CH approach, section 4 presents the system model, section 5 discusses the numerical results obtained, and finally paper is concluded with future scope in section 6.

II. Related Work

Several mechanisms of CH election exist with an objective to endow with efficient mobile computing in terms of stable routing in the MANET system [12, 13]. Some mechanisms favor not changing the CH to reduce the signaling overhead involved in the process, which also makes the elected MU usage of its own resources higher [14]. Other mechanism assigns the CH based on the highest MU ID as in the Linked Cluster Algorithm, LCA [15]. However, this selection process burdens the MU due to its ID. CH can become bottleneck and lead to propagating congestion. One option is to elect CH for a defined duration and then all MUs have a chance to be a CH [3]. This mechanism keeps the CH load within one MU for the CH duration budget, while it provides a balance of responsibilities for MUs within the cluster. Also, MU with a high mobility rate may not get the chance to become a CH if its mobility rate is higher than the duration of CH rotation. But transition and the duration budget contribute greatly to overhead. Mobility is one of the most important challenges of MANETs, and it is the main factor that would change network topology. A good electing

About-[1] Professor Deptt. of Comp. Science & Applications, Kurukshetra University, Kurukshetra, Haryana, India (e-mail;pksuritf25@yahoo.com)
About-[2] Asstt. Prof., M.M. Inst. of Comp. Tech. & B. Mgmt., M.M.University, Mullana, Haryana, India(e-mail; kavitatane@gmail.com)

CH does not move very quickly, because when the clusterhead changes fast, the MUs may be moved out of a cluster and are joined to another existing cluster and thus resulting in reducing the stability of network. Hence, CH election mechanisms consider relative MU mobility to ensure routing path availability [16, 17], however, causing an added signaling overload and causing the elected CH to pay the higher resource utilization penalty. We can conclude from the existing research that several tradeoffs exist for the elected CH and the other cluster MUs. Firstly, the CH has to bear higher resource utilization such as power, which may deplete its battery sooner than other MUs in the cluster. In addition, possibly causing more delay for its own application routing due to the competition with the routing for other MUs. Secondly, despite fair share responsibility of CH role, it is possible that heavy burst of traffic takes place causing some CHs to use maximum resources, while others encounter low traffic bursts resulting in minimum resource use. Thirdly, the fair share or load balancing technique [3], might result in a CH that will not provide the optimal path for routing, or yet a link breakage. Plus non CH are privileged as they don‗t pay a routing penalty and have resources dedicated for own usage only. Therefore, there is no one common CH election mechanism that is best for MANET systems, without some hurting tradeoffs. The Zone Routing Protocol (ZRP) [18] provides a hybrid approach between proactive routing which produces added routing control messages in the network due to keeping up to date routes, and reactive routing which adds delays due to path discovery and floods the network for route determination. ZRP divides the network into overlapping zones, while clustering can have distinct, non overlapping clusters. In ZRP, Proactive routing is used within the zone, and reactive routing is used outside the zone, instead of using one type of routing for the whole network. In addition, [18, 19] suggest that hybrid approach is suited for large networks, enhances the system efficiency, but adds more complexity. Each MU has a routing zone within a radius of n hops. All MUs with exactly n hops are called peripheral MUs, and the ones with less than n are called interior MUs. This process is repeated for all MUs in the network. A lookup in the MU‗s routing table helps in deciding if the destination MU is within the zone resulting in proactive routing. Otherwise, the destination is outside the zone, and reactive routing is used which triggers a routing request. As a result of a routing response, one of the peripheral MUs will be used as an exit route from the zone to the destination. While, if clustering is applied, the same elected CH is used for routing outside the cluster without triggering any route discovery to the destination. As discussed above, the main focus of the existing work focuses on an election of single CH for a cluster. Even though this minimizes the overall signaling overhead in the cluster, but it mainly can make the central CH a bottleneck.

A. Challenges And Issues In Clustering

Despite the tremendous potentials and its numerous advantages MANET pose various challenges to research community. This section briefly summarizes some of the major challenges faced while clustering in such network [12-15].

B. Heterogeneous Network

In most cases MANET is heterogeneous consisting of MUs with different energy levels. Some MUs are less energy constrained than others. Usually the fraction of MUs which are less energy constrained is small. In such scenario, the less energy constraint MU are chosen as CH of the cluster and the energy constrained MUs are the member MUs of the cluster. The problem arises in such network when the network is deployed randomly and all cluster heads are concentrated in some particular part of the network resulting in unbalanced cluster formation and also making some portion of the network unreachable. Also if the resulting distribution of the CHs is uniform and if we use multi hop communication, the MUs which are close to the CH are under a heavy load as all the traffic is routed from different areas of the network to the CH is via the neighbors‘ of the CH. This will cause rapid extinction of the MUs in the neighborhood of the CHs resulting in gaps near the CHs, decreasing of the network size and increasing the network energy consumption. Heterogeneous MANET require careful management of the clusters in order to avoid the problems resulting from unbalanced CH distribution as well as to ensure that the energy consumption across the network is uniform.

C. Network Scalability

In MANET new MUs comes in the vicinity of the current network. The clustering scheme should be able to adapt to changes in the topology of the network. The key point in designing cluster management schemes should be if the algorithm is local and dynamic it will be easy for it to adapt to topology changes.

D. Uniform Energy Consumption

Clustering schemes should ensure that energy dissipation across the network should be balanced and the CH should be rotated in order to balance the network energy consumption.

E. Multihop or Single Hop Communication

The communication model that MANET uses is multi hop. Since energy consumption in wireless systems is directly proportional to the square of the distance, most of the routing algorithms use multi hop communication model since it is more energy efficient in terms of energy consumption however, with multi hop communication the MUs which are closer to the CH are under heavy traffic and can create gaps near the CH when their energy terminates.

F. Cluster Dynamics

Cluster dynamics means how the different parameters of the cluster are determined for example, the number of clusters in a particular network. In some cases the number might be reassigned and in some cases it is dynamic. The CH performs the function of compression as well as

The distance between the CHs is a major issue. It can be dynamic or can be set in accordance with some minimum value. In case of dynamic, there is a possibility of forming unbalanced clusters. While limiting it by some pre-assigned, minimum distance can be effective in some cases but this is an open research issue. Also CH selection can either be centralized or decentralized which both have advantages and disadvantages. The number of clusters might be fixed or dynamic. Fixed number of clusters cause less overhead in that the network will not have to repeatedly go through the set up phase in which clusters are formed. In terms of scalability it is poor.

III. Multi – CH Approach.

The existing clustering approach encourages election of one CH [20, 21]. The proposed work enhanced the architecture to use multiple CHs and distributes the load of the single CH amongst multiple CHs in the same cluster. The proposed mechanism does not mandate a specific CH election process. Any of the prior work [9, 10] can be used to select the CHs for a cluster. By distributing the load, a single CH does not have to bear all the added responsibility of being the central point for routing in a cluster. Therefore, we believe this approach provides a more fair solution of sharing inter-cluster routing responsibilities for a cluster. In addition, other mechanism can be applied to switch the responsibility of a CH to another MU, such as in [3]. In the case of one CH per cluster, a link breakage caused by the failure of the CH isolates all cluster MUs from communicating to/from outside the cluster. However, our approach reduces the link breakage to be only in the direction towards a path where the failed CH forwards the data. Therefore, the reliability of routing in the MANET system is increased. We explore the certain benefits of having multiple sinks in the network as follows:

Energy efficiency: In MANET, long routing path lengths from MU located at the cluster borders to the CH are observed. Adding extra CH to the cluster decreases the average path length between a MU and the CH due to shorter geographic distance between them. Therefore, the number of hops that a packet has to travel to reach a CH gets smaller. Since each traveled hop means the data packet consumes some energy at the visiting MU, traveling fewer hops results in consuming lesser energy.

Avoiding congestion near a CH: Using multiple CHs can also relieve the traffic congestion problem associated with a single-CH system.

Avoiding single point of failure: A single-CH is not robust against failure of the CH or the MU around the CH. Multi-CH are therefore more resilient to MU failures. However, deploying many CHs does not solve the problem directly and evenly. It is essential to distribute cluster load among CHs and choose an optimal route(s) between MU and the corresponding CH. transmission of data.

IV. System Model

We have used glomosim [22] simulator, running IEEE 802.11 to prove our contribution. Our MANET system consists of four distinct non-overlapping clusters with a physical terrain of 1500 meters by 1500 meters as shown in Fig. 1. For the same cluster, we ran simulation experiments with one CH, and compared its performance results with tests using 3 CHs. Each CH has an independent queue for packets destined for the neighboring clusters for which a particular CH is meant. During the simulation, we maintained the same CHs in both cases (single, multiple CHs), since changing the CH was irrelevant to what we are proving. Our traffic type has Constant Bit Rate, (CBR), and File Transfer Protocol, (FTP), traffic. The same traffic load was run for both cases (single, 3 CHs). The selected traffic load was chosen based on tests that allowed sufficient utilization of the channel.

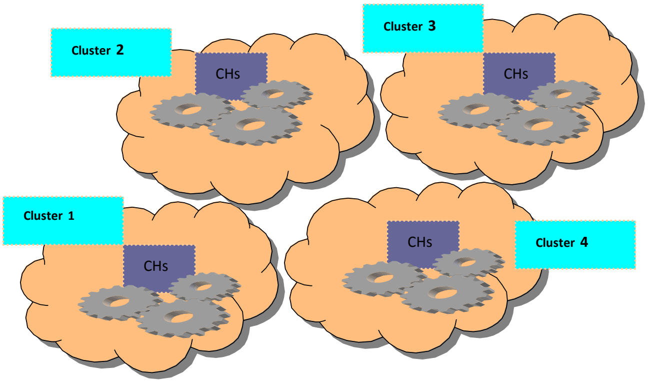


Fig.1. Multi-CH Simulation Setup

In this model Cluster 4 operates as a cluster with one CH and with many CHs. The remaining clusters operate with one CH. This work can be expanded by incrementing the number of CHs in a cluster such that it has one CH per neighboring cluster. Our traffic included FTP traffic generated between MUs in all clusters in the MANET system. The FTP sessions where established in both directions. In addition, CBR traffic was generated in both directions between MUs in cluster 4, and clusters 1, and 2. In order to focus on the objective of distributing the CH load, we setup static routes in our MANET system. Routing from cluster 4 to cluster 2 was done via the intermediate cluster 1/cluster 3, and vice versa. Therefore, since there are 3 neighboring clusters to cluster 4, the system allowed for the use of 3 CHs, one for routing to/from each neighboring cluster.

V. Numerical Results

Our simulation focused on the cumulative averaged throughput and response time. Fig. 2 shows the percentage of increase in throughput when running multiple CHs over using one CH. In all cases, the throughput increased for the multiple CHs case. For the small simulation time of 1000S and with the traffic load used, the increase was only about

18% since the system was lightly loaded as a result of a short simulation time. Therefore, one CH operated well since the channel was not well utilized. Our peak results show that at 7000S of simulation time, we reached a maximum throughput improvement as this case indicates the channel utilization was at its optimal condition. Therefore, for the longer simulation times, beyond what we concluded as optimal, the throughput decreased due to the added traffic on the channel.

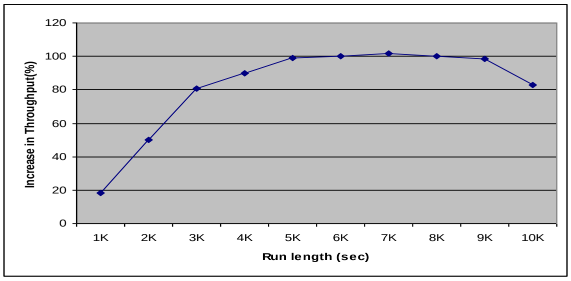


Fig.2. Run length (sec) VS Throughput Improvement (%)

The optimal case of 7000S proves the advantage of distributing the load to multiple CHs, we have gained about 101% improvement in throughput. Our results are explained by the simple queuing theory model:

$\rho = \lambda / \mu$ (1)

where, $\boldsymbol{\rho}$ is the traffic intensity, $\boldsymbol{\lambda}$ is the traffic arrival rate and $\boldsymbol{\mu}$ is the service rate at each CH with queue length QLI (*k*,*l*) with *k* as no. of packets and *l* as no. of CHs per cluster. Eq.1 indicates that ρ increases if the λ increases while μ remains at the same rate. In addition, the overall averaged cumulative response time, increases if a constant service rate is maintained, while the traffic arrival rate increases. Our simulation showed that the response time remained constant when using one single CH, and multiple CHs of about 0.5. The traffic rate in the system is given by Box Muller transformation (Eq. 2) with given σ=1 and μ=0 and rand1, rand2 as samples from U (0, 1).

$s = (-2 \text{ Log } (\text{rand1})^{1/2} \text{ Cos } (2\pi. \text{ rand2})$ (2)

The traffic rate is increased as indicated by the throughput increase due to the multiple CHs, while maintaining the same response time. Normally, if the arrival rate increases while maintaining the same service rate, then the response time should increase accordingly. Therefore, we can conclude that, by maintaining the same response time, the added traffic rate due to an increase in service rate results in constant system utilization. In our topology, we increased the number of CHs to 3. However, our throughput is about doubled as shown in Fig. 2. We should expect by the distribution of work to 3 CHs, and by having the same averaged delay for the MANET system, a 3 fold increase in throughput since the service rate has tripled. However, we only gained double the throughput due to cumulative increase in overall overhead due to the added traffic rate by having multiple queues, one for each CH. In addition, as the traffic arrival rate increased due to having the 3 CHs, the service rate also increased, resulting in the same utilization rate for the MANET system. We ran additional test to validate the traffic rate at our selected simulation time of 7000S. The tests were run with one CH and multiple CHs for cluster 4. The throughput results are presented in Fig. 3. The results show the percentage of increase in the averaged cumulative throughput for running multiple CHs over one CH. We ran test at 4 traffic rates: High, medium (half of the high), low traffic rate (half of the medium) and at much lower traffic rate than the low traffic rate which we called very low rate traffic.

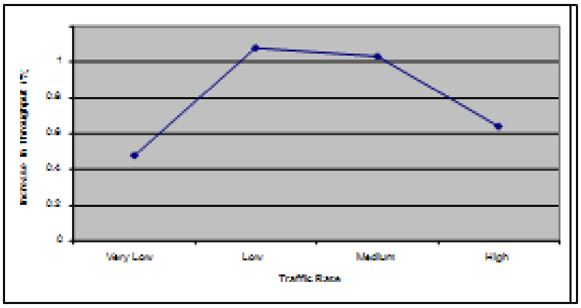


Fig.3. Throughput Improvement (%) VS Traffic Rates

We have noticed, as shown in Fig. 3, the percentage of throughput improvement for the very low was only nearly 50%. This is attributed to the low channel utilization by the low traffic rate. At the high traffic rate we have shown a reduced improvement in throughput due to traffic overload and multi queue overhead in the MANET system. This traffic overload was created by the higher arrival rate due to the added sessions. However, at medium traffic rate, we obtained about the same level of throughput improvement as our optimal selected rate. We conclude that at these rates we obtained system stability with the offered traffic and service rates with many CH. Therefore, the results shown in Fig. 3 validate the selected traffic for our results above

VI. CONCLUSIONS AND FUTURE WORK

Our contribution proves that one CH per cluster does not provide for a maximized throughput of the MANET system due to the added responsibility for the one CH. Using multiple CHs (with independent queue) per cluster distributes the load among multiple MUs which enables simultaneous and shared responsibility of inter cluster routing among multiple MUs. It is an interesting finding to note that the increase in throughput due to the added CHs is proportional to the number of CHs. Beat with the number equal to the neighboring clusters. Depending on the topology and traffic pattern, if all CHs are simultaneously used to route traffic, the rate of throughput increase fails to be the multiplier of the original throughput when using one CH due to overhead of maintaining multiple CHs in a cluster. It is suggested to do further research when having all clusters employing multiple CHs, one per neighboring clusters. Also one expansion of the system model is to take one common queue and dispensing the packet to the idle CH irrespective of the neighboring cluster route. It is expected that the throughput will increase at a very high rate as MANET is blessed with multi hop communication and minimizing the idle time of CHs will lead to balancing the overhead caused by their existence.

VII. REFERENCES

1) Buss, D. 2005, ―Technology and design challenges for mobile communication and computing products," in proceedings of the 2005 International Symposium on Low Power Electronics and Design (San Diego, CA, USA, Aug. 08-10, 2005). ISLPED '05. ACM, New York, NY.
2) S. Sivavakeesar, and G. Pavlou,"Stable clustering through mobility prediction for large-scale multihop intelligent ad hoc networks," in proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'04), Georgia, USA, Mar. 2004, vol. 3, 1488-1493.
3) Amis, and R. Prakash, ―Load- Balancing Clusters in Wireless Ad Hoc Networks," in proceedings of the 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology (ASSET'00), pp 25, Mar. 2000.
4) M. Gerla, and J. Tsai, ―Multicluster, Mobile, Multimedia Radio Network," ACM Journal on Wireless Networks, vol. 1, no. 3, pp 255-265, 1995.
5) Nocetti, J. S. Gonzalez, and I. Stojmenovic, ―Connectivity based k-hop clustering in wireless networks," Telecommunication Systems Journal, vol. 22, no 1-4, pp. 205-220, 2003.
6) Arboleda C., L. M. and Nasser, N, ―Cluster-based routing protocol for mobile sensor networks," in proceedings of the 3rd international Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (Waterloo, Ontario, Canada, August 07 - 09, 2006). QShine '06, vol. 191. ACM, New York, NY, 24.
7) Akkaya K., Younis M., "A survey on routing protocols for wireless sensor networks", Elsevier Ad Hoc Network Journal, vol.3, no. 3, pp. 325-349, 2005.
8) Cardei, I., Varadarajan, S., Pavan, A., Graba, L., Cardei, M., and Min, M, ―Resource management for ad-hoc wireless networks with cluster organization," Cluster Computing, vol.7, no.1, pp. 91-103, Jan. 2004.
9) Wang, S., Pan, H., Yan, K., and Lo, Y, ―A unified framework for cluster manager election and clustering mechanism in mobile ad hoc networks," Comput. Stand. Interfaces, vol. 30, no. 5, pp. 329-338, Jul. 2008.
10) V. S. Anitha , M. P. Sebastian, ―Scenario-based diameter-bounded algorithm for cluster creation and management in mobile ad hoc networks," in proceedings of the 2009 13th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications, pp.97-104, Oct. 25-28, 2009
11) Spohn, M. A. and Garcia-Luna-Aceves, J. J., ―Bounded-distance multi-clusterhead formation in wireless ad hoc networks," Ad Hoc Networks vol. 5, no. 4, pp. pp. 504-530, May. 2007.
12) Khac Tiep Mai , Dongkun Shin , Hyunseung Choo, ―Toward stable clustering in mobile ad hoc networks," in proceedings of the 23rd International Conference on Information Networking, pp.308-310, Jan. 21-24, 2009, Chiang Mai, Thailand.
13) X. Hong, M. Gerlo, Y. Yi, K. Xu, and T. J. Kwon, ―Scalable ad hoc routing in large, dense wireless networks using clustering and landmarks," in proceedings of the IEEE International Conference on Communications (ICC'02), vol. 25, no. 1, pp. 3179-3185, Apr. 2002.
14) ER, I. I. and Seah, W. K., ―Clustering overhead and convergence time analysis of the mobility-based multi-hop clustering algorithm for mobile ad hoc networks," in proceedings of the 11th international Conference on Parallel and Distributed Systems - Workshops ICPADS. IEEE Computer Society, vol. 02, pp. 130-134 Washington, DC, Jul. 20 - 22, 2005.
15) Jane Y. Yu and Peter H.J. Chong, ―A survey of clustering schemes for mobile ad hoc networks" IEEE Commun. Survey & Tutorial, vol 7 no. 1, pp. 32-48, Mar. 2005.
16) C. R. Lin and M. Gerla, "Adaptive clustering for mobile wireless networks," IEEE JSAC, vol. 15, pp. 1265-75, Sept. 1997.
17) A,McDonald, T. F. Znati, ―A mobility based framework for adaptive clustering in wireless ad hoc networks," IEEE JSAC, vol. 17, no. 8, pp.1466- 1486, Aug. 1999.
18) Z. J. Haas, and M. R. Perlman, ―The performance of query control schemes for the zone routing protocol," in proceedings of ACM Sigcomm'98, vol. 28, no. 4, pp 167 – 177, Oct. 1998.
19) P.Y. Chen, and A.L. Liestman, ―Zonal algorithm for clustering an hoc networks," International Journal of Foundations of Computer Science, in a special issue dedicated to Wireless Networks and Mobile Computing, vol. 14, no. 2, pp. 305-322, Apr. 2003.
20) Zang, C. and Tao, C., ―A multi-hop cluster based routing protocol for MANET," in proceedings of the 2009 First IEEE international Conference on information Science and Engineering (December 26 - 28, 2009). ICISE. IEEE Computer Society, Washington, DC, pp. 2465-2468, 2009.
21) Wang, C., Yu, Y., Xu, Y., Ma, M., and Diao, S., ―A multi-hop clustering protocol for MANETs," in proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing (Beijing, China, September 24 - 26, 2009). IEEE Press, Piscataway, NJ, pp. 3038-3041, 2009
22) .Web site for glomosim simulator, http://pcl.cs.ucla.edu/projects/glomosim/

Corporate Data Obesity: 50 Percent Redundant

Hae Kyung Rhee

GJCST Computing Classification
H.2.m, K.6.4

***Abstract*-In this essay, we report what we have observed with regard to status quo of corporate information systems in real world from our experiences of twenty years of data management practices. It is considered to be serious in that data are too conveniently and frequently replicated to make information systems improperly behave in terms of their quality standards including response time. Average ratio of data replication in a site is astonishingly judged to be more than 50 percent of a whole corporate database. It is in reality about 65 percent in average to our knowledge. Presenting this paper to academia has been motivated by our strong belief and evidence that most of the redundancy can effectively and systemically be removed from the very start of information system development. We also noted that field workers including database administrators in corporate environment tend to think data part of IS and program part of IS mixed together from the start of IS design and popularity of this tendency eventually caused a lot of entanglement that could hardly be dealt with later by themselves. We therefore present a couple of mandates that must be respected in order not to get involved in such a perplexity**

Keywords-Corporate Data Obesity, Data Redundancy, Enterprise Data Map.

I. Concept of Obesity

It is not unusual to think that if a person is weighed more than about 20 percent of what needs to maintain for fitness then he or she is considered to be over-weighted. This is what we understand with regard to concept of obesity. It is no different for data in corporate environment. It will be astounding to recognize that the degree of data obesity in corporate is far more than 20 percent. It is in fact 65 percent in average for some dozens of large enterprises we have observed in depth for the past twenty years. To be exact in terms of terminology, the unit of obesity we mean is data attribute. For example, if there is a customer data and it is comprised of c-name and c-address, c-name and c-address are the data attributes. So, in case c-name appears more than once in a corporate database, it is called redundant or replicated. Although the reports on data abundance in corporate environment have been made in the literature, as far as we know, only the issue of data deluge [Cukier2010, KaBoZe2010] has been dealt with a couple of times in order to emphasize world-wide phenomenon of rapidity in increase of data in terms of volume. The issue of data obesity is new in the world-wide communities of database research and management information systems research. In this sense, it is almost impossible to find any past work in the literature made with regard to this issue. Note that the concept of data obesity is essentially irrelevant to data volume. Although introduction of some upper-level data stores like data warehouses (DW) or data marts (DM) other than the lower-level operational data stores (ODS) in corporate environment certainly contributes to abundance of data, DWs and DMs are out of scope in this essay. If we stick only to ODSs, we could observe that a lot of obesity is already there in corporate environment.

Note that, in a fairly large corporate such as General Electric or Samsung Electronics, there are approximately 15,000-to-20,000 data attributes in their database. Notice also that the level of redundancy in data attribute is not exactly the same as the level of redundancy in data volume. However, to make it comparatively simple to have some idea about redundancy in terms of data volume, since a lot of people in field work prefer this way of understanding, when we happen to hear that database size of some company is, for instance, 100 terabytes, it is legitimate or reasonable to think that the company in reality has a database of approximately 35-to-50 TBs. So, in case 50-to-65 TBs of data can be totally eliminated from the corporate database and this elimination does never affect harm the normal operation of the database at all. Redundancy demands a huge cost in terms of waste in storage and belatedness in response to database queries. Note that even 1 TB of data amounts to piling A4 size papers up about 100 kilometers high.

Redundancy or replication gives some illusion that it could contribute to enhancement of response time, but on the other hand things can get messy if we consider consistency of data. The quality of answers to data queries could be always in question, since making all the replica copies to have the same value usually takes a substantial amount of time due to non-automatic processes of such data value propagation. Manual propagation by considerate programming nevertheless unfortunately incurs unforced human errors and there is no guarantee for data consistency at all across a corporate database. Once an inconsistent value of data happens to be used to reply the queries, trust of information system would unbelievably collapse. Issue of mistrust would then raise the question of integrity with regard to a whole information system.

Therefore, limiting the occasions of data replication to be minimal is necessary whenever it is possible. Unless the rate of data redundancy is substantially reduced, say to about 15 percent by means of wary design from the outset of IS development, data normalization theories [YuJa2008] that have been esteemed almost over the past thirty years turn out to be ―useless" at all in real world. To our knowledge

About- Associate professor at Dept. of Computer Game & Information in Yong-In Songdam college.
(telephone:82-31-330-9234 e-mail;leehk@ysc.ac.kr)

reduction comes quite before some tabular form of data begins to emerge in the process of IS development and that is just where we start to lay out job descriptions, in non-technical term. We will get back to this later in this essay after discussion with regard to how people in IT field are insensitive to the issue of redundancy.

II. Unnecessary redundancy

an arena where data is represented in a form of table or relation, in expertise terminology, the concept of keys like primary key and foreign key is technically inevitable. Basically, if a particular key of table, say A, dubbed its primary key, is duplicated in another table, say B, as a part or component of key of B, that key is denoted as a foreign key in B, as it has been imported or borrowed from other table, which is A. This clarifies that origin of the key is from A, not B. This way of designating and incorporating such externality of key will bring IS about 15 percent of data redundancy contained intrinsically, which is technically unavoidable if we stick to the tabular representation of data. This portion of redundancy can be called redundancy of necessity. So, if data obesity ratio is said to be 65 percent, it is true that about 45 percent of the entire data is therefore classified to be unnecessary or superfluous in their nature.

Whether to remove this much of unnecessary redundancy or unwanted replication is up to decision of an individual data manager, but unless removal of them is done the information system would definitely be hampered or suffered by lack of consistency and further by eventual slowness in response time. Note that, normally in the database queries of any corporate, about half of them are update requests and the other half are retrieval requests. If this reality of read-write ratio, i.e. 0.5, is ignored, we are soon tempted to allow data duplication by assuming that reads are much more frequent than writes, and subsequently a fatal disaster would then be experienced sooner or later due mainly to data inconsistency dilemma.

The payoff for burden of upholding this unnecessary redundancy is really enormous. Usually, it would be about five times more costly than the case where the level of redundancy is minimally enforced. So, it is going to be 10 million dollars versus 50 million dollars when so called next generation, i.e. enhanced version, of information system is to be developed. As the degree of data redundancy increases, data consistency tasks among operational databases exponentially as well increase in proportion to the amount of increase in data redundancy. Note that there is inevitably redundancy between the lowest-level database and its upper-level data warehouses, since data in database are in principle shoveled upward to its data warehouses in the process of generating data warehouses. It is also a natural consequence that another layer of redundancy is unavoidable between data warehouses and their upper-level data marts.

In case data redundancy is existent, it is not difficult to find many of duplication are intrinsically semantic. Syntactic duplication is easy to find out, but it is almost impossible to determine whether any data is a semantic derivative of some other data. This semantic data duplicity is the major malice to make corporate database incurably obese. So, it is necessary to remove syntactic duplication, but it is exceedingly more crucial not to forge any possibility of semantic duplicity from the very outset of IS development. It really is almost impossible to check semantic equivalence, even periodically, once an information system is in operation day to day.

III. De-normalization—panacea or deadly homepathy?

It is really unfortunate that we have never seen any data table or relation that even follows the rule of well-known first normal form (1NF) in real world corporate databases. So, sometimes it is ridiculed that real world databases only contain tables of non-normal form or zero normal form, since they have properties significantly inferior than 1NF in terms of data quality such as the degree of data redundancy and dependability of non-key data attributes to key attributes. The beauty of table normalization or table standardization by applying 1NF, 2NF, 3NF or Boyce-Codd NF is that whenever there is a data redundancy in a table then it is possible to remove it by decomposing or splitting the table into two.

In corporate IT field unfortunately a term ―de-normalization” [JoJA2007] has gained so much popularity in a sense that field managers usually do not have a time to pay attention to and understand the theories behind normalization. They at first pretend to understand and use them, but in reality they sooner or later totally forget about them. By far, we are very unfortunate that we have never seen any database administrator who really does understand the basic difference between 1NF and 2NF. The reality is that they keep never trying or studying to grasp the meaning and benefit of making tables normalized and keep feigning to have started with 1NF initially for IS development and to proceed forward to make tables in up to 3NF and all of sudden for the sake of performance they inevitably and eventually come to resort to 1NF again. But this could be a sort of fictional story and hence never true at all, since they always had failed to tell us what the intrinsic difference between 1NF and 3NF is. A number of experiments [KSLM2008] already have shown that having tables in 3NF performs always better than 2NF or 1NF and that 3NF is considered to be quite optimal even in cases where seven-way table joins are conducted. Note that 7-way join means that combining seven different tables, each fairly large in our experiments, at the same time.

The real problem with IT field managers and even database administrators is that they hardly understand even what the 1NF is. Note that in any data-related literature for the past forty years of history, notion of ―de-normalization” has never been introduced, but they pretty much fond of taking that jargon just in order to forget about normalization stuff and to wish to let themselves totally unaware of any impending issues related to data consistency. They seem to be soon relieved to hear by someone else that normalization could always be compromised for the reason of performance. To our knowledge, they are misled by mainly outside IT consultants who have never been trained enough in basic knowledge in database. So, it is actually a very demanding burden to make them understand what the normalization theories are all about.

However, this is not too bad if we know that having tables even in 3NF could contribute to reduce the degree of data redundancy by at most about 5 percent, which is not too much. Consequently, the contribution of normalization would be only minor. But then, where is the majority of contribution come from? It comes much prior to the formulation of tables. In order to realize this, we have to know what and where the origin of data essentially is in corporate environment. Where is the place where redundancy really starts to build? It is at the very beginning of business processes, not where the normalization theories are just about to be applied. Wouldn't it be curious that where are all the data that are to be appeared eventually in tables come from?

IV. Necessity of Business Processes Description

Let us turn our attention to how business processes are described so that field workers can communicate each other later on. They will certainly be in a form of business processes description or job description. So, the transformation of job descriptions into data tables might take a couple of interim stages, since descriptions themselves have a format different from table and there is no direct, straightforward method that can map the descriptions into tables. Then, how is job description comprised of? In it, there could appear data entity like employee or department which has fixed values for data attributes it is comprised of.

For example, a data entity ‗employee' might consist of data attributes ‗address' and ‗social security number' and their values are normally fixed, i.e., not changed over time. In case in job description there is a description statement like ―Anemployee sells a machine.", data entities ‗employee' and ‗machine' will have such fixed values, while on the other hand data entity ‗sell' is different in that the values that data attributes of ‗sell' like selling date or selling volume vary, i.e., changed each time the action or behavior ‗sell' is performed. So, action entities are at the focal point in terms of creating different data values in the database. It can be considered that the source entity of action ‗sell' is ‗employee' and its destination entity is ‗machine'. This way of writing job descriptions by taking action-oriented approach or behavior-oriented approach [KDLM2007] is straightforward. It could be fairly easy to understand for employees who have a mission of writing a description for jobs they actually perform.

Efforts to make job descriptions to be free from data redundancy are essential and valuable to check whether there is redundancy of any sort for each particular action. This means the action ‗sell' above appears at most only once in job descriptions of whole business processes of a corporate. It is judged to be improper or abnormal if the action ‗sell' appears more than once in entire job descriptions of the corporate. This kind of effort in reducing or removing actions redundancy has no relationship in what is known to be crucial like 1NF, 2NF or 3NF, as emphasized in the literature. But removal effort with regard to redundancy in data attributes directly associated with actions is far more important than the removal of redundancy in tables at a later stage of database creation. If the removal effort is not sufficiently done, redundancy thus retained intentionally or unintentionally would then automatically be transferred intact to tables at the instance of table creation.

From the perspective of who or what is in charge of dynamically creating data in corporate environment, it is fair to admit that behaviors, rather than fixed entities, play the major role of such creation. Fixed entities that are always expressed as nouns in description statements like ‗employee' and ‗department' normally generate only static data attributes and thus said to be only at the outskirt in data-creating activities. In this sense, it is meaningful if we preferably write job descriptions in a way of behavior-by-behavior. Each behavior then has a responsibility for creating only meaningful data attributes. In case a behavior does not contribute to generate certain attributes, it has no value of existence to be independent or stand alone. This means that in that case it is reasonable to place that behavior to be subsumed by some other behavior that is directly relevant and superior to it.

V. Behavior-Oriented Job Descriptions

As we have observed over the past 20 years, the unit of resources that is assigned to an employee is normally a job. Definition of jobs has been in a sense pretty much well established in corporate. For example, we could count the number of jobs in a corporate without much difficulty. To our experience, a mid-size corporate has about 500 to 1,000 jobs and to perform those jobs it normally requires to maintain the number of employees of about twice as much as the number of jobs, since it is a usual practice to assign two persons to a single job in order to prepare for emergencies of just-in-case. So far, we have seen a number of corporate that have about 500 jobs and 1,000 employees in real world. This might be a kind of standard for mi-size corporate.

We were able to observe from our experience that each job in average could be comprised of some 20-to-30 actions or behaviors in case data-creating actions are only taken into account in job descriptions. So, if there are 500 different jobs in a corporate, then it means that there are about 10,000-to-15,000 behaviors altogether in that company. With no redundancy in actions, those some 10,000 behaviors must be unique in that they do not incur redundancy of any types so that each of them must appear once and at most once throughout the entire corporate database.

VI. Enterprise Data Map

These behaviors are in a sense interconnected each other in a way that each data-creating action has one fixed entity on its left and one more fixed entity on its right. If we denote a interconnection would look like a type of ‗E—B—E'. So, behavior by B and a fixed entity by E, then the web of those the whole picture would look something like a rectangular type that would allow data accesses or data retrievals in either direction, clockwise or counter-clockwise, as depicted in arrows in Fig.1.

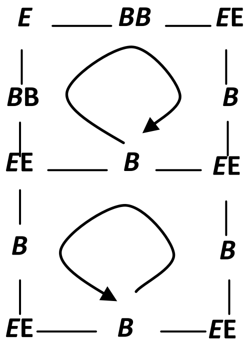


Fig. 1. Rectangular Path Formed in Enterprise Data Map, where B Denotes Behavior and E Denotes Entity

Rectangularity guarantees balance in response time in either direction of access, while if otherwise skewed case to one particular direction could induce degradation in response time. Although there are only seven actions in this picture, we could get a whole diagram that contains some 10,000 behaviors if we keep extending the picture by adding more behaviors to it. The entire picture of connection without allowing isolation of any picture fragment could be called an enterprise data map [Moon2004].

With this EDM, we are able to judge or realize where the origin of a particular data attribute is and how it flows throughout the entire data access paths already obtained and depicted in EDM. With EDM, it is very easy to find out visually where are data redundancies if there are any. As a diagram, one EDM can depict about 20 pages of A3-size in case font size of 5 is used. Drawing would be automatic if we use a software drawing tool such as ERwin [JoJB2007]. The EDM of such many pages would then easily fit into the wall of CEO‘s or CIO‘s office. Or it could also be displayed on CFO‘s office in case he is interested in figuring out how is the flow of all the data directly related to financial status quo of his company. Unfortunately, at the moment only a few corporate experienced the value of obtaining and maintaining the EDM, but we advocate that its use would significantly benefit many aspects of information system. We advocate that utilization of EDM would thereafter be plentiful according to your perspectives of looking at it.

VII. Separation of Data from Program

It is needless to say that EDM is the must to be secured and kept as an asset prior to the programming of information system. We emphasize that any programming effort must be deferred until the finalization of EDM. EDM in this sense is the blueprint for any design like, for instance, building or road. To our knowledge, EDM is definitely the blueprint for information system prior to any programming effort. What we emphasize is that data itself is essentially data in that programming must begin to take place only after the data formulation has been made to sure to be completely wrapped up. Data-first programming-later approach is crucial for the success of information system. If data stuff and programming stuff are mixed together from the start of information system development, chaotic situations would duly be encountered in determining that whether an impending problem at issue is originally from data part or programming part. We emphasize that any data cannot be represented or expressed or substituted in a way of any programming means.

Note that if somebody happened to introduce a data _whether-a-student-is-registered-or-not‘, then it is in fact a disguise as a data in that it essentially has a sort of algorithmic logic in that data. Presuming that a data like _registration date‘ could reside somewhere else in the database already, _whether-or-not‘ type of decision could then be definitely dealt with some conditional statements like _if‘ in programming. Separation of data from programming must be strictly obeyed in a sense that, without separation, a bunch of semantic redundancy like this sort of disguise could later be insidiously come into the information system. If it seems that this way of algorithmic logic is certainly in a data, then it is not real data, since only the raw data is privileged to be called as data. Anything impure in a way of generating artifacts is not called the real data. For example, if data C is from the result of addition of raw data A and raw data B, then C is not in principle treated as data. Note that in the lowest infrastructural level database of corporate only such raw data are entitled to reside. Anything else must be deported to reside somewhere else like data warehouses.

VIII. Conclusion

In sum, there are two major mandates that have to obey to make information systems free from data obesity. The first one is that efforts for removing data redundancy should be enforced from the start of information system development, which is from the starting point of securing job descriptions. The latter one is the strict separation of data arena and programming arena in developing information systems. Questions like whether this belongs to data or programs are better to be raised as frequently as possible in order not to bring any chance of confusion about which comes before and which comes after or later. To our knowledge, the degree of data obesity is guaranteed to be tolerated within at most 20 percent if these two mandates are strictly obeyed.

Removal of another 5 percent of data redundancy is later possible if we conduct a certain set of technical details. The well-known data table normalization or data table decomposition theories come into play for this further removal. So, the benefit accrued from the data redundancy removal efforts by application of normalization theories is considered to be far less than we get from the efforts made at the stage of job description, which is about 30-to-45 percent of removal in data redundancy in an entire corporate database. It is adding one more flower to a beauty itself already seized if the normalization theories are applied to make tables best fit with minimal redundancy in them, but we certainly might have no regret at all when they happen to be not applied for some reason under the premise that data redundancy of all sort has already been sorted out and managed to be ruled out prior to table formulation.

The adage ―Trying to start with guarantees almost half-way done already" still prevails in the world of information system development and making IS fit or well-being in any situation or environment comes true when we immersed to think in this manner. Consequently, the earlier we preoccupied with the trial of data redundancy removal, the better the outcome of information systems in terms of performance, clarity, transparency and promptness in response time.

IX. References

1) [Cukier2010] Cukier, K. (2010, Feb. 25). Data, Data Everywhere. A Special Report on Managing Information, The Economist. Retrieved May 1, 2010,from http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443.html

2) [KaBoZe2010] D. Katz, M. Bommarito, & J. Zelner (2010, March 1). The Data Deluge. The Economist print edition.
3) [JoJB2007] J. Jones & E. Johnson (2007). Building and Maintaining A Database from An ER Model. White Papers : Computer Associates.
4) [Moon2004] S. Moon (2004). Data Architecture, Hyung-Seol Publishing Company.
5) [KDLM2007] N. Kim, D. Lee and S. Moon (2007). Behavior-Inductive Data Modeling for Enterprise Information Systems. Journal of Computer Information Systems, Vol. 48, No. 1, 105-116.
6) [KSLM2008] N. Kim, S. Lee & S. Moon (2008). Formalized Entity Extraction Methodology for Changeable Business Requirements. Journal of Information Science and Engineering, Vol. 24, No. 3, 649-671.
7) [YuJa2008] C. Yu & H. V. Jagadish (2008). XML Schema Refinement through Redundancy Detection and Normalization. VLDB Journal, Vol. 17, 203-223.

Web Mining: A Key enabler for Distance Education

D.Santhi Jeslet[1] Dr . K.Thangadurai[2]

GJCST Computing Classification
H.2.8, J.1, H.3.5,H.5.3

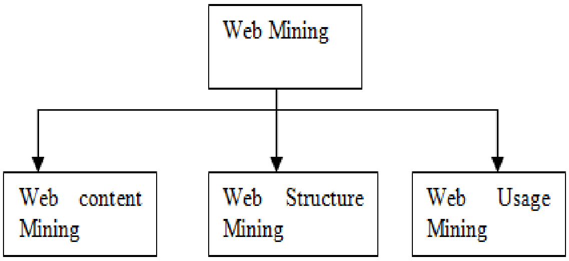
***Abstract*-This paper deals with introduction of one of the application of data mining which is know as web mining. It discuss about various categories of web mining. It also deals with the application of web mining in distance education and describes the possibilities of application. In this fast world everyone wants to be educated by acquire huge knowledge in a short duration. They do not want to spend some fixed time for their education. Whenever a person is free they can learn and gain the knowledge.**

Keywords-Data mining, web mining, distance education

I. Introduction

Now-a-days many organizations accumulate huge amount of data. This leads to swell the size of the database as the time passes. Traditional database queries access a database using SQL queries. The output of this could be data from database that satisfy the query. This output cannot give any novice information or correlation among the data. So we need a technique that finds the hidden information from data collection in a database community which is of large size. This technique is called the ―Data Mining”. It discovers valid, novel, potentially useful new correlation and new trends from the large amount of data. Data mining uses pattern recognition techniques, statistical and mathematical techniques for its discovery.

In the recent trend, lots of databases are available in the web. Not only the database, many valuable informations are also available in WWW. So the search area for any information has become very vast. Web mining is an application of data mining which uses the data mining techniques to automatically discover and extract information from Web documents/services. It can also be applied to semi-structured or unstructured data like free-form text. Web mining activities can be divided into three categories: content mining, structure mining, and usage mining. The taxonomy of Web mining is depicted in the figure.



1. Web Content Mining: It is the process of discovering useful information from the web which may be in the form of text, images, audio and video. For the discovery it uses the techniques of Artificial Intelligence (AI), Database and most specifically Data Mining (DM).
2. Web Structure Mining: It helps to derive knowledge of interconnection of documents, hyperlinks and their relationships. It uses graph theory to analyze the node and connection structuring of a web site.
3. Web Usage Mining: It is also called as web logs mining. This helps to judge about the usage of a web page. It uses computer network concepts, artificial intelligence and database.

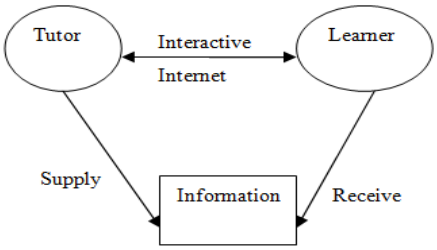
II. Objectives of distance education

In the last few decades education has undergone many changes. Class room teaching is needed for face to face education which comprises of class room, presence (physical) of some learners and a teacher/tutor. Here teacher/tutor plays a vital role. But by the introduction of distance education, the interactions between the tutor and the learner have been very much reduced. Even the interaction between the learners has become almost zero.

The main aim of distance education is to make the society to acquire more knowledge irrespective of the place where they are. Those who do not want to stick on to the rules of regular education system, prefer to earn knowledge through distance education. It also encourages working people to attain their learning goals.

III. How web helps in distance education?

The communication between the tutor and the learner can be enhanced by the introduction of distance education through web. Here learners work individually at their own place, with the help of some study materials i.e. system, computer program and internet. Time and space limitation of education disappears. Tutors interact with the student and the learner interacts with the tutor via internet. The tutor supply information and learner gets it.



Since many softwares are very simple and user friendly, no need to get special training for working with computer. Power of computers makes student to improve their ability.

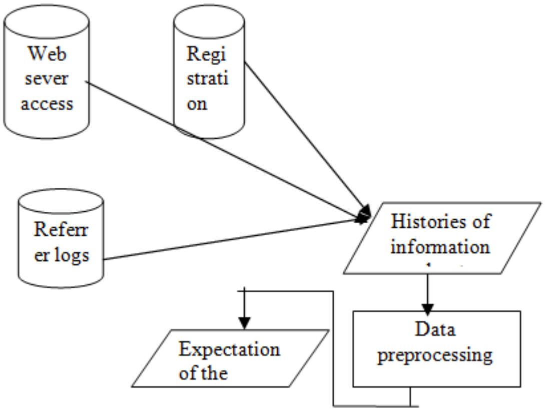
About-[1] Department of Computer Science, M.G.R.College,Hosur,TN, India (e-mail: santhi.jeslet@rediffmail.com)
About-[2]Department of Computer Science, Government Arts College(Men) Krishnagiri- 635 001, India

Role of tutor is entirely changed. Tutor communicates and leads the course of their learning path. Learners will be grouped. They learn from each other and they also assess each other. It allows the learners to apply their knowledge in different situations and to solve practical problems according to the feedback of their own action. These changes in educational system have developed constructivism. Constructivism means learners involve actively constructing meaningful knowledge through experience.

IV. Application of Web Mining in distance education

Organization that is responsible for distance education collect huge volume of data, which are generated automatically by web servers and collected in the server access logs. They also collect information from learner (referrer) logs which contain information about the referring pages for each page and also from user registration. Through this an organization can get idea about thinking styles, learns their expectations and also about the web site structure. This helps to improve the efficiency of the web site that is responsible for improving the knowledge of the learners.

Before gathering histories using mining algorithms, number of data preprocessing issues such as data cleaning has to be performed. The major preprocessing task is data cleaning. This is used for removing irrelevant information in the server log.



The extracted access histories of each individual learner are representing the physical layout of web sites with web page and hyperlinks between the pages. Once user access histories have been identified, perform web page traversal path analysis for customized education and web page association for virtual knowledge structures.

By using different path analysis such as graph representation we can determine most frequently traversal patterns form the physical layout of a web site. Path analysis is performed from two points of view: aggregate and individual path. Aggregate path includes the process of clustering the registered learners. The web site database has the registered learner‘s details. This can be segmented by one of the clustering techniques to discover learners with similar characteristics. By using this we can determine most frequently visited paths of learners. Individual path helps to determine a set of frequently visited web pages accessed by a learner during their visits to the server.

By discovering such aggregate and individual paths for learner in distance education helps in the development of effective customized education. Associations and correlation among web pages can be discovered using association rules. This guides to discover the correlations among references to various web pages available on the server by a learner or learners. Based on this the tutor can also judge the standard of the learner.

V. Conclusion

Web mining in distance education provides a lot of open teaching resources, so that people can teach and learn anytime and anywhere. It helps the organization that is responsible for distance education to discover the learner‘s access habit and the study interest. It guides the teacher to adjust his/her teaching techniques and the speed of teaching depending on the learner‘s knowledge. So web mining technology is a key enabler of distance education.

VI. References

1) Youtian QU, Lili ZHONG, Huilai ZOU, Chanonan WANG. ―Research About The Application Of Web Mining In Distance Education Platform. Scalable Computing And Communication, Eighth International Conference On Embedded Computing,2009.SCALCOM EMBEDDEDCOM‘09 International Conference On Digital Object Identifier
2) WANG Jian And LI Zhuo-Ling. Research And Realization Of Long-Distance Education Platform Based On Web Mining, Computational Intelligence And Software Engineering 2009, Cise 2009, International Conference On Digital Object Identifier
3) Sung Ho Ha, Sung Min Bae, Sang Chan Park. Web Mining For Distance Education, Management Of Innovation And Technology,2000,ICMIT 2000, Proceeding Of The 2000 IEEE Conference On Volume 2, Digital Object Identifier
4) Zhang Yuanyuan, Mo Quian. Research Of Constructivism Remote Education Based On Web Mining , Education Technology And Computer Science 2009, ETCS‘09, First International Conference On Volume 2, Digital Object Identifier
5) Margaret H. Dunham And S. Sridhar. Data Mining: Introductory And Advanced Topics
6) Pieter Adriaans And Dolf Zantinge. Data Mining
7) Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques

Optimized Remote Network Using Specified Factors As Key Performance Indices

John S.N[1] Okonigene R.E[2] Akinade B.A[3]

Chukwu I.S[3]

GJCST Computing Classification
C.2.1, C.2.5

***Abstract*-This paper discuss the implementation of an optimized remote network, using latency, bandwidth and packet drop rate as key performance indicator (KPI) to measure network performance and quality of service (QoS). We compared the network performance characteristics derived on the Wide Area Network (WAN) when using Fiber, VSAT and Point-to-Point VPN across the internet respectively as the network infrastructure. Network performance variables are measured across various links (VAST, Fiber and VPN across the internet) and the corresponding statistical data is analyzed and used as base-line for the optimization of a corporate network performance. The qualities of service offered on the network before and after optimization are analyzed and use to determine the level of improvement on the network performance achieved.**

Keywords-Key performance indicator, optimized remote network, latency, bandwidth, WAN, VSAT.

I. Introduction

Most network users often attribute the problem of slow network and poor quality of service to lack of sufficient bandwidth, which is not generally correct. Sometimes, poor network performance can be traced to network congestion, high packet drop rate, chatty protocols and high latency [1] among others. This paper uses the technique of network base lining to obtain the best combination of network metrics that can enhance the performance of network resources up to maximum data flow energy (MDFE) which allows maximum amount of data to be sent in the fastest amount of time using the optimum bandwidth capacity [2]. We assume that the Server and client processing time are minimal relative to the total time it takes to complete a transaction. Hence, it attributes the cause of service transaction delays to WAN delay. It try to find out the causes of poor quality of service across the WAN and makes recommendation or how to implement efficient remote network with better quality of service (QoS) [3]. In the methodology, three sets of parallel links (Fiber, VSAT and Point-to-Point VPN across the internet) of equal bandwidth are set up between two geographically separate locations. Files of different size were sent between the locations across each link respectively. The key performance indicators (latency, bandwidth and packet drop rate) [4, 5, 6] were recorded using standard monitoring tools to monitor each of the experiment performed. Graphical analysis of the data obtain from the link performance were used as the bases for the conclusion made in this paper using latency, bandwidth and packet drop rate as key performance indicator for network performance.

II. Network Performance Criteria

A network can be rated as performing when end-users are able to access applications and carry out given task without undue perceived delay, error or irritation. The primary measure of user perceived performances are availability and completion time. It is important to identify whether utilization factors, collision rate or bandwidth congestion are responsible for network problems [7]. In general, the performance of a computer network can be divided into three sections for easy analysis and trouble-shooting:

- The performance of the application,
- The performance of the servers,
- The performance of the Network infrastructures.

Based on end-user perception of the network, we can also view the network performance in terms of service oriented and efficiency oriented as shown in the Fig. 1.

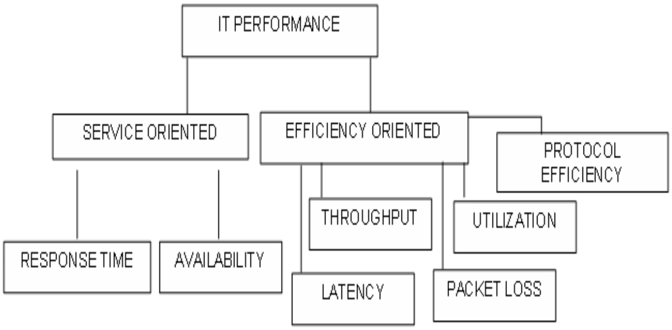


Fig:1. Block diagram of IT performance

It is noted that, service oriented performance measures how well an application provides service to the customer, whereas efficiency oriented performance measure how much of available channel resource are actually used to provide end-user request. This tend to measure how much of available channel resources are being wasted due to inefficiencies inherent in the communication channel.

About-[1] Department of Electrical and Information Engineering, Covenant University, Ota, Nigeria (e-mail; johnsam8@hotmail.com)
About-[2] Department of Electrical & Electronics Engineering, Ambrose Alli University, Ekpoma, Nigeria (e-mail; robokonigene@yahoo.com)
About-[3] Department of Electrical and Electronics Engineering, University of Lagos Akoka, Yaba, Lagos Nigeria (email;bayonleakinade@yahoo.com) (e-mail; sunnymentus@yahoo.com)

III. METHODOLOGY

The performance of a wide area network can be verified by studying the effect of network contribution to transaction time (NCTT) on the network [3].

In a high performance network, TCP packets are transferred across the WAN with minimal delay (low latency) within the optimum load limit. When the network becomes overloaded, congestion sets in and TCP packets are drop and consequently re-transmitted which adds to the total time required to complete a transaction in a busy network [8]. Network contribution to transaction time is the sum of the round-trip times necessary to complete a given transaction type, plus the time for recovery from any lost packets during the transaction [3]. The network contribution to transaction time can be calculated as:

$$NCTT = E * RTT + L * RTO$$

where, E – number of round-trip exchange necessary to complete the transaction,

RTT – round-trip time for packet transfer,

L – number of round-trips exchanges that experience packet loss,

RTO – retransmission time-out

The number of losses experienced in the course of a transaction depends on round-trip packet loss probability, *p*.

For a two-ways traffic path, loss probability is given by:

$$P = P_{RTT} = 1 - \{(1 - P_{oneway}) * (1 - P_{otherway})\}$$

If each round-trip exchange, takes A_i attempt to complete successfully, and the total attempts to complete a transaction given as:

$$A = \sum_{i=1}^{E} A_i$$, then

$$\Pr ob(A_i = a) = p^{a-1}(1-p)$$

Expected value of *A* is given by:

$$E = \{A\} = E_x \sum_{a=1}^{\infty} axp^{a-1}(1-p) = \frac{E(1-p)}{p} \sum_{a=1}^{\infty} axp^{a}$$

this converge as:

$$E\{A\} = \frac{E}{1-p} \text{ for } 0 < \mathrm{p} < 1$$

A is equal to the constant *E* plus a random number of losses *L*, so $E\{A\} = E + E\{L\}$

$$E\{L\} = \frac{E}{1-p} - E = E(\frac{p}{1-p})$$, and the average

$$NCTT = E * RTT + [E\{L\} * RTO]$$

Note that the probability distribution of NCTT is a set of discrete values [11] at

(E x RTT),

{(E x RTT) + (1 x RTO)},

{E x RTT) + (2 x RTO)},

The performance of the WAN and remote network can also be viewed in terms of its effective throughput.

Throughput is the quality of error-free data that can be transmitted over a specified unit of time [9].

$$Throughput = \frac{Bandwidth}{TotalLatensy}, bps$$

$$\text{Also, } Thoughput = \frac{MSS}{RTT} * \frac{1}{\sqrt{p}}, \text{ bps}$$

where,

MSS – Maximum segment size (fixed for each internet path, typically 60 bytes)

RTT – Round trip time (as measured by TCP)

P – Packet loss rate (%)

The efficiency of the WAN link can be calculated from statistical data on the link utilization, where Utilization (U) [7] is the percentage of total channel capacity currently being consumed by aggregate traffic.

$$Utilizatio\,n = \frac{Traffic}{Channel\ capaciy} * 100$$

Also,

$$Utilizatio\,n = \frac{[(Data\ sent + data\ received) \times 8}{Link\ speed \times sample\ time} * 100$$

Further more, in this research, three point-to-point WAN link were setup between two separate locations A and B using three different WAN technologies, namely:

(i) 128/256Kbps leased fiber line

(ii) 128/256Kbps point-to-point VPN across the public internet.

(iii) 128/256Kbps VSAT link

The key performance indicators (KPI) metrics for the research were Latency, Bandwidth and Packet Drop Rate. The following approach methods were used to obtain the required performance characteristics of the various WAN technologies adopted:

(a)Files of various sizes were sent from Host A to Host B across the different WAN links.

(b)These KPI values were measured and recorded for different remote network infrastructure in use (Fiber, VSAT, Point-to-Point VPN across the internet with bandwidth of 128/256 kbps respectively)

(c) The performance statistic values obtained in both cases were plotted in graphical form and analyzed.

(d) Recommendation for error correction and performance improvement were made

(e) Conclusion was drawn based on the result obtained from the key performance indices.

The alternative WAN links between two remote locations shown in Fig. 1, were routed to Host A and Host B using different connection links (Fiber, VSAT, P2P VPN) to measure the KPI of the network.

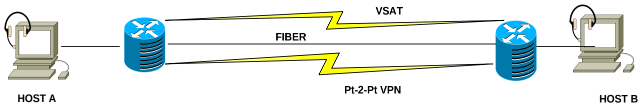


Fig. 1. Schematic diagram of alternative WAN links between two remote locations

The Table 1, shows the result of the throughput obtained from the remote link of the WAN with different Packet Drop Rate of the links.

Table 1. Throughput result of a network as affected by both the latency and the packet drop rate

LATENCY (ms)	THROUGHPUT						
	TP1(KBPS) 0.01% PDR	TP2(KBPS) 0.05% PDR	TP3(KBPS) 0.10% PDR	TP4(KBPS) 0.50% PDR	TP5(KBPS) 1.00% PDR	TP6(KBPS) 2.00% PDR	TP7(KBPS) 3.00% PDR
9	1822.22	814.95	576.29	257.70	182.22	128.85	105.20
30	546.67	244.48	172.89	77.31	54.67	38.66	31.56
60	273.33	122.24	86.44	38.66	27.33	19.33	15.78
90	182.22	81.50	57.63	25.77	18.22	12.86	10.52
120	136.67	61.12	43.22	19.33	13.67	9.66	7.87
150	109.33	48.90	34.58	15.46	10.93	7.73	6.32
300	54.67	24.45	17.29	7.73	5.47	3.87	3.16
500	32.80	14.70	10.37	4.64	3.28	2.32	1.90
800	20.50	9.17	6.48	2.90	2.05	1.45	1.18
1000	16.40	7.34	5.19	2.32	1.64	1.16	0.95

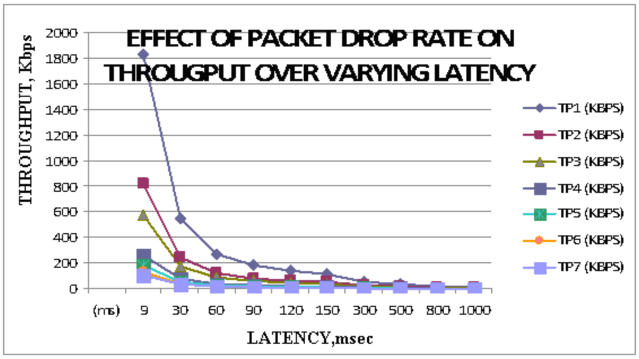


Fig. 3. Graph of throughput against latency for different packet drop rates

The Fig. 3, shows the effect of packet drop rate on the network throughput over different latency. The throughput of a network is affected by both the latency and the packet drop rate of the link where an increase in latency decreases the network throughput performance. Similarly, the throughput also decreases as the packet drop rate increases which might put the network quality of service to network degradation. Analysis of the achieved result indicates that, the best quality of service will be obtained by using a link whose latency is between 1 – 30 milliseconds and packet drop rate of 0.01% or less. Such latency can only be achieved using Fiber or radio link where packets are propagated at the speed of light with very low bit-error-rate.

The worst quality of service occurred when latency is between 800 – 1000 milliseconds and the packet drop rate stands at 3% or more.

The link latency of 800 milliseconds and above is usually associated, with VSAT link because of its technological limitation caused by distance along the propagation path between two locations via the orbital satellite.

However, VSAT links could still be used for none delay-sensitive application if there are no packet loss. The situation becomes worse when increasing packet drop rate is associated with VSAT links. For a Point-to-Point virtual private network (VPN) across the public internet with average latency of 250 milliseconds, most real-time and data-based applications performance is considered favorable. However, Point-to-Point VPN is always associated with higher packet drop rate than VSAT or Fiber links because of the large number of hop and routing protocols across the part from source to destination. This is even worse when considering a two-way traffic situation usually experienced in real life scenario.

IV. Improvement in Quality of Service

The improvement in quality of service (QoS) can be seen by comparing the network throughput of the Fiber, VPN, and VSAT link of a network. If we assume a minimal packet loss for all the three infrastructures: latency of 850ms for VSAT, Point-to-Point VPN across the internet at 260ms and Fiber link of 25ms.

Throughput for VSAT gives 0.6168Mbps that of the VPN across the internet gives 2.016Mbps and the throughout for fiber gives 20.97 Mbps. By replacing the VSAT infrastructure with Fiber Optic link, the following improvement in QoS would be achieved.

Hence the improvement in QoS gives

$$\frac{20.97-0.6168}{0.6168}*100 \text{ => } 3300\%$$

Similarly, replacing the VPN with Fiber optic link would be achieved with an improvement in quality of service QoS as follows:

$$\frac{2.016-0.6168}{0.6168}*100 \text{ => } 530\%$$

V. Conclusion

The Key Performance Indices of network services (packet drop rate, latency and throughput) affects the network performance as one the factors goes out of the optimized range value obtained in the research work.

Under perfect conditions (assuming minimal percent of packet loss), the use of WAN link with low latency, and use of optimized bandwidth would significantly enhance the quality of service (QoS) experienced by a remote network user over a WAN link.

VI. Reference

1) Bilal Haider, M. Zafrullah,M.K. Islam, —Radio frequency optimization & quality of service evaluation in operational GSM network," Proceedings of the world Congress on Engineering and Computer Science 2009, Page 1, Volume 1. WCECS, Oct 20-22, 2009, San Francisco USA.
2) Daniel Nassar, —Network Performace Baselinning," Publisher MTP, 201 West, 103rd Street, Indianapolis, IN46290 USA, 2002.
3) Network contribution to transaction time. ITU-T Recommendation G.1040 ITU-T Study Group 12 under the ITU-T Recommendation A.8 procedure (2005-2008).
4) —Effect of network latency on load sharing in distributed systems," Journal of parallel and distributed computing volume 66, issue 6 Inc Orlando, FL USA (June 2006).
5) Jorg Widmer, Catherine Boutremans, Jean-Yves Le Boudec: End-to end congestion control for TCP –friendly flows with variables packet size. Publisher: ACM, 2004.
6) Gregory W. Cermak, —Multimedia Quality as a function of bandwidth, packet loss and latency," International Journal of Speech Technology. Publisher: Springer Netherlands Issue: Volume 8, Number 3, 2005.
7) Michael Lemm, —How to improve WAN application performance," 2007 http://ezinearticles.com.
8) Lai King Tee, —Packet error rate and Latency requirements for a mobile wireless access system in an IP network," Vehicular Technology Conference, Pages 249 -253, 2007.
9) J.Scott Haugdahl, —Network Analysis and Troubleshooting," Publisher: Addison –Wesley, 2003.

Analysis of the Routing Protocols in Real Time Transmission: A Comparative Study

IKram Ud Din[1] Saeed Mahfooz[2]

Muhammad Adnan[3]

GJCST Computing Classification
C.2.2

***Abstract*-During routing, different routing protocols are used at the routers to route real time data (voice and video) to its destination. These protocols perform well under different circumstances. This paper is about to evaluate the performance of RIP, OSPF, IGRP, and EIGRP for the parameters: packets dropping, traffic received, End-to-End delay, and variation in delay (jitter). Simulations have been done in OPNET for evaluating these routing protocols against each parameter. The results have been shown in the graphs which show that IGRP performs the best in packets dropping, traffic received, and End-to-End delay as compared to its other companions (RIP, OSPF, and EIGRP), while in case of jitter, RIP performs well comparatively.**

Keywords-Routing, Protocol, Delay, Packet Loss, Jitter

I. Introduction

A protocol is a set of rules that reveals how computer systems communicate with each other across networks. A protocol also functions as the common medium by which different hosts, applications, or systems communicate. The data messages are exchanged when computers communicate with one another. Examples of messages are sending or receiving e-mail, establishing a connection to a remote machine, and transferring files and data. There are two classes of protocols at the network layer, i.e., routed and routing protocols. The transportation of data across a network is the responsibility of the routed protocols, and routing protocols permit routers to appropriately direct data from one place to another. In other words, protocols that transfer data packets from one host to another across router(s) are routed protocols, and to exchange routing information, routers use routing protocols. IP is considered as a routed protocol while routing protocols are: i). Routing Information Protocol (RIP), ii). Interior Gateway Routing Protocol (IGRP), iii). Open Shortest Path First (OSPF), and iv). Enhanced Interior Gateway Routing Protocol (EIGRP), etc. To forward data packets, the Internet Protocol (IP) uses routing table. RIP uses hop count to determine the path and distance to any link in the internetwork. In case of multiple paths to a destination, RIP selects the path that has fewest hops. The only routing metric RIP uses is hop count; therefore, it does not necessarily opt for the fastest path to a destination [1]. IGRP is developed to address the problems associated with routing in large networks that are beyond the scope of RIP.

IGRP can select the fastest path based on the bandwidth, delay, reliability and load. By default, it uses only bandwidth and delay metrics. To allow the network to scale, IGRP also has a much higher maximum hop-count limit than RIP. OSPF was developed by the Internet Engineering Task Force (IETF) in 1988. OSPF shares routing information between routers belonging to the same autonomous system. It was developed to address the needs of scalable, large internetworks that RIP could not. EIGRP is an advanced version of IGRP that provides superior operating efficiency such as lower overhead bandwidth and faster convergence [1].

As we are examining the video and voice packets during video conferencing and voice packet transmission in this paper, therefore a short introduction of those protocols must also be inevitable that are used for the transmission of these packets. In video conferencing, Real Time Transport Protocol (RTP) is used for carrying out video packets, and for session establishment between the two systems, either H.323 or SIP is used. RTP provides end-to-end network transport functions premeditated for real time applications such as video and voice. Those functions comprise payload-type identification, time stamping, delivery monitoring and sequence numbering [2].

Voice over Internet Protocol (VoIP) is a means of compressing voice using a standardized codec, then encapsulating the results within IP for transport over data networks. For establishing and transporting VoIP traffic, H.323 is a standard protocol [3].

The H.323 standard has been developed by the ITU-T for vendors and equipment manufacturers who provide VoIP service. It was originally developed for multimedia conferencing on LANs, but was later extended to VoIP. The 1st and 2nd versions of H.323 were released in 1996 and 1998, respectively. Currently, its version 4 is under consideration.Session Initiation Protocol (SIP) is the Internet Engineering Task Force (IETF) standard for multimedia or voice session establishment over the Internet. It was proposed as a standard in February 1999. SIP: a detailed protocol that stipulates the commands and responses to set up and tear-down calls. It also details features such as proxy, security, and transport (TCP or UDP) services. SIP describes end-to-end call signaling between devices. SIP defines, as the name implies, how the session is established between two IP nodes with or without media [2].

About-[1,2,3] Department of Computer Science, University of Peshawar, Pakistan
(e-mail[1] ;ikramuddin205@yahoo.com)
(e-mail; saeedmahfooz@yahoo.com)
(e-mail; adnan5283@yahoo.com)

The goal of this study is to measure the performance of throughput, packet loss, jitter, and delay in real time transmission. The simulations have been done in OPNET, because OPNET has originally been developed for network simulation, and it is fully usable as an ample simulation tool with higher investment. OPNET provides a complete development environment for the specification, simulation and performance analysis of communication networks [4], [5], [6]. OPNET must be able to simulate different network devices and various kinds of transmission lines, and display such information as packet end-to-end delay, delay variation (jitter), and packet loss in the network. The main purpose is to analyze how the network having speech activity. The voice quality can be characterized by two measurements: i) delay of the signal, and ii) distortion of the signal. The delay disturbs the interactivity, while distortion reduces the legibility [7]. Many factors such as a heavy load in the network that creates higher traffic, may contribute to the congestion of network interface [8]. Therefore, this research is important to be managed in order to measure and predict data transfers in real time applications. The remaining paper is structured as: Section 2 describes the work done in the evaluation of routing protocols. Section 3 illustrates the working environment for the implementation of these protocols. Section 4 explains the OPNET simulations of the mentioned protocols. Section 5 concludes our work, and references are given in section 6.

II. Related work

Privacy and security become necessary requirements for Voice over IP (VoIP) communications that need security services such as integrity, confidentiality, non-replay, non-repudiation, and authentication. Quality of Service (QoS) of the voice is affected by jitter, delay, and packet loss [9].

Normally, telecommunication network consists of routers which optimize the packets' transmission. Practically, a packet is transmitted through a number of paths from one router to another. The selection of path is based on routing tables' information usually received according to routing protocol. A routing protocol is one that provides techniques facilitating a router to build a routing table. It also shares routing information with other neighboring routers.

When a router is switched off, the packets passing through that router is passed to another router. This operation is known as "routing protocol convergence". Packets are possibly to be lost during a routing protocol convergence [10].

Networks like the Internet are renowned today. Such networks consist of routers, switches and hubs, communication media, and firewalls. Servers and clients are usually interconnected by networks. During communication through the Internet, there may be many possible routing paths and many routers between a source and destination. When packets arrive at a router, the router decides as to the next hop in a path to the destination. For making this decision, many algorithms are used, such as RIP, OSPF, IGRP, and EIGRP, etc. The RIP and OSPF try to route the packets to a destination via the path consisting fewest number of nodes (routers). The IGRP and EIGRP attempt to route the packets based on shortest path, shortest delays, and greatest bandwidth factors.

The invention of Curtis et al [11] makes routing decisions. In their invention, a best path is determined according to an IGRP, EIGRP, OSPF, BGP or other routing task that can provide multiple routing paths. A first variety of routers in the best routing path is determined.

Their invention also makes decision for routing a received packet. If the first variety of routers had a noise level, the packet is forwarded to a next router in the best routing path. If not, then according to said IGRP, EIGRP, OSPF, BGP, or the other routing function in a second routing path is determined [11].

A network facilitates the delivery of packets from a source to destination. This delivery is possible through routers. Packets have destination addresses that let routers to determine how to route the data packets. A router has a routing table which stores network-topology information. With the help of network-topology information, the router forwards packets to the destination. A routing protocol consists of methods to select the best path and exchange topology information. There are two main classes of routing protocols: distance vector routing protocols, e.g. RIP and IGRP, and link-state routing protocols, e.g. OSPF. For enterprise networks, OSPF is often preferred [12], [13].

To exchange service availability and network reachability information, router implements one or more routing protocols. In a specific implementation, the border router implements RIP, OSPF, IGRP, EIGRP, or BGP [14].

Routing protocols accept network state information and then on the basis of such accepted information, update network topology information. Routing protocols also distribute the network state information. Path generation and forwarding information generation are also duties of the routing protocols [15], [16].

III. Working environment

When a node wants to transmit real time applications (video or voice) over IP then it must have to pass through a router. For transmission of real time applications, real time transport protocol (RTP) is used and the session is established between two remote stations through session initiation protocol (SIP) or H.323. Except, these real time transmission protocols, some routing protocols are also used which route the real time applications to its destination. These are: RIP, OSPF, IGRP and EIGRP.

Consider the following scenario having two servers i.e. VoIP and video, and two clients which are: VoIP and video client. The distribution of the servers and clients are at two different location, i.e., servers are located at site Lahore (in this case) and the clients at the other site (say Karachi).

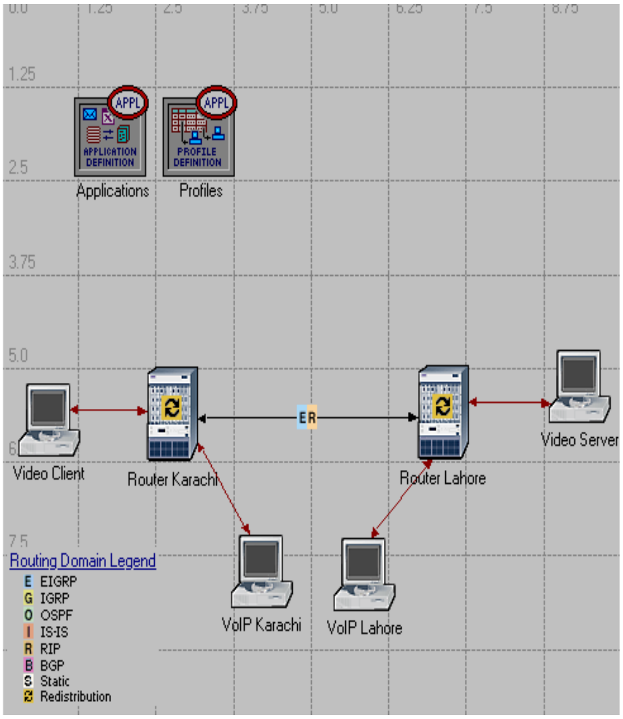


Fig. 1: structure of the network

A. *IP Packet/Traffic Dropping*

When a router or switch is unable to receive incoming data packets at a given time, is called Packet loss/drop. The real time applications (video or voice) are drastically degraded by packet loss [17].

B. *Video/Voice Traffic Receiving*

Video/voice traffic is the total number of audio and video packets received during video conferencing or other type of real time communication (e.g., IP telephony).

C. *End-to-End delay*

End-to-end delay depends on the end-to-end data paths/signal paths, the payload size of the packets, and the CODEC. Delay is the latency; one-way or round-trip, encounter when data packets are transmitted from one place to another. In order to maintain the expected voice quality for Voice over IP (VoIP), the roundtrip delay must remain within almost 120 milliseconds. [17].

D. *Variation in Delay (Jitter)*

In computer networks, the term jitter means variations in delay of packets received. Jitter is an essential quality of service (QoS) factor in evaluation of network performance. It is one of the significant issues in packet based network for real time applications [18]. The variation of interpacket delay or jitter is one of the principal factors that disturbs voice quality [19]. Jitter plays a vital role for the measurement of the Quality of Service (QoS) of real time applications. The effect of end-to-end delay, packet loss, and jitter can be heard as: The calling party says, ―Hello Sir, how are you?" With end-to-end delay, the called party hears,…...Hello Sir, how are you? With packet loss, the called party hears, He.lo….r, w are you? With jitter, the called party hears, Hello…Sir, how....are… you? [2].

IV. SIMULATION RESULTS

In this section, a scenario was tested in which the delay, packet loss, and jitter were examined.

Figure 2 shows the number of IP packets dropped per second. Figure 3 illustrates the traffic received during video conferencing. The voice traffic received is shown in figure 4. The end-to-end delay in voice packets is given in figure 5, while variation in delay or jitter is clear from figure 6.

A. *Performance Evaluation*

The number of packets dropped is given in figure 2; in which the less number of packets is lost when IGRP is implemented at the routers. While a huge amount of packets is dropped if OSPF works as a routing protocol. IGRP also works well in case of receiving video and voice packets, given in figure 3 and 4, respectively. The end-to-end delay and variation in delay (jitter) in voice traffic is shown in figure 5 and 6, respectively, in which IGRP is also the best protocol. In the given figures, the X-axis shows the amount of time and the Y-axis shows the number of packets in figure 2, 3, and 4, and in figure 5 and 6, it shows the value of jitter and delay.

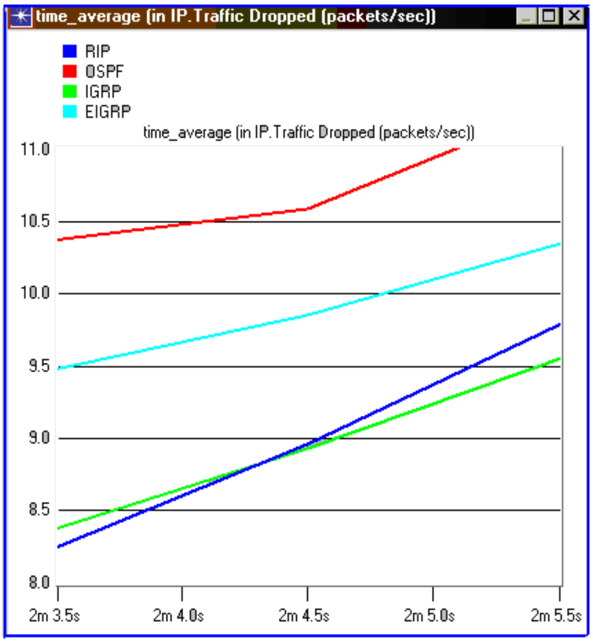


Fig. 2: Number of packets dropped per second

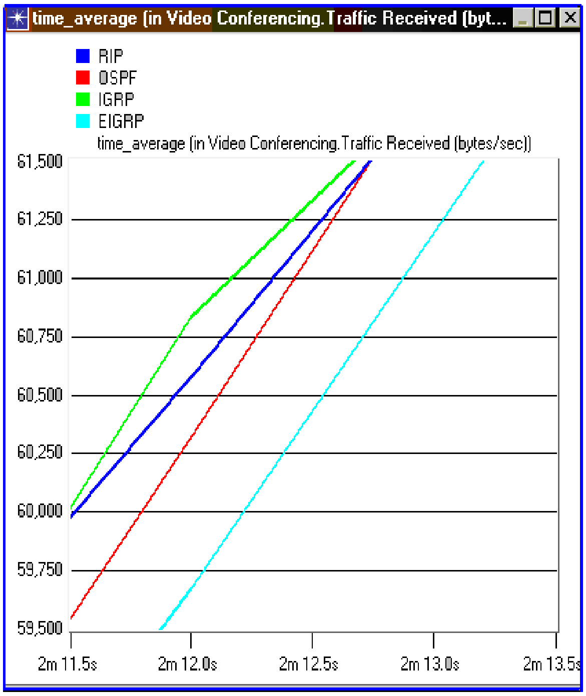


Fig. 3: video traffic received per second

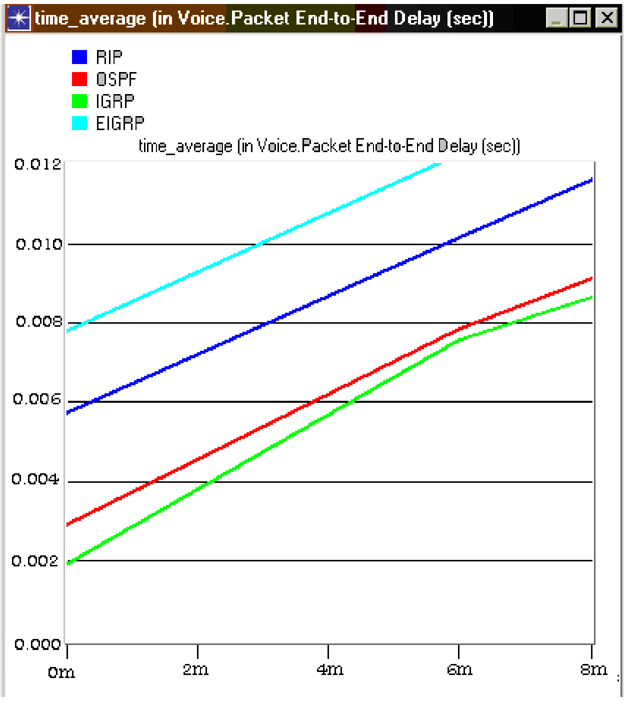


Fig. 5: End-to-End Delay in voice Packets

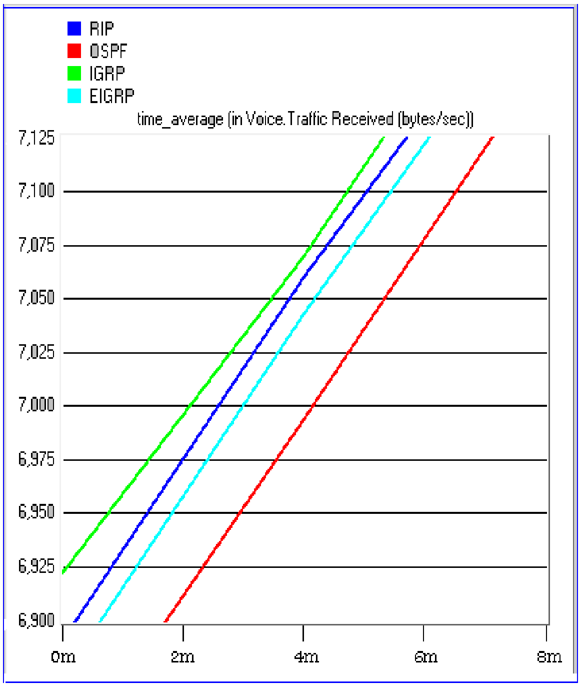


Figure 4: voice traffic received per second

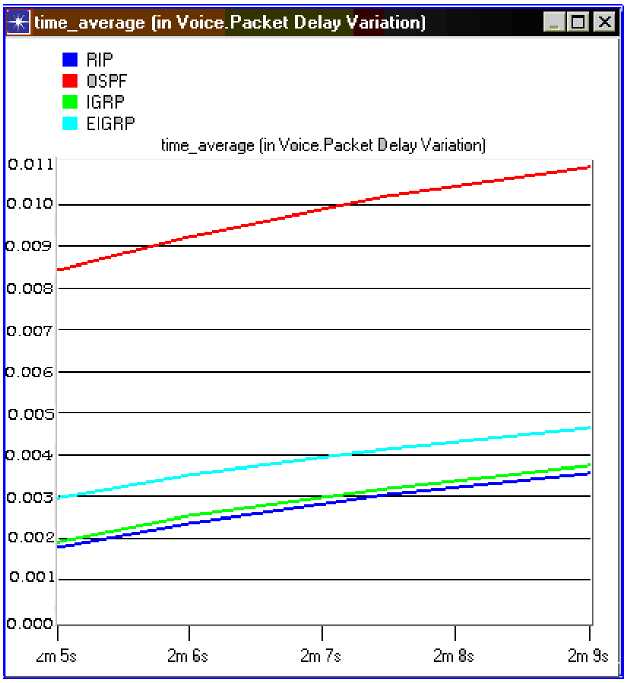


Fig. 6: Jitter in Voice Packets

V. Conclusion

The size of today's networks has been growing quickly and support complicated applications, e.g., video conferencing and voice messages. Quality transmission is demand of the time. This needs some good results producing routing protocols at the routers. The work done in this paper analyzes the available routing protocols: RIP, OSPF, IGRP and EIGRP for packets dropping, traffic received, End-to-End delay, and variation in delay (jitter). Our work is based on OPNET simulation for each of these parameters. The study presents a comprehensive result for each protocol against the parameters: packets dropping, traffic received, End-to-End delay, and variation in delay (jitter) one by one. IGRP performs well in packets dropping, traffic received, and End-to-End delay as compared to its other companions (RIP, OSPF, and EIGRP), while in case of jitter; RIP performs a bit well than IGRP.

VI. References

1) Cisco Systems, I., Cisco Networking Academy Program CCNA 1 and 2 Companion Guide Third Edition. 2003.
2) Cisco Systems, I., Cisco Voice Over IP. Student Guide, ed. V. 4.2. 2004.
3) Shufang Wu, M.R., Riadul Mannan, and Ljiljana Trajkovic, OPNET Implementation of Megaco/H.248 Protocol. 2003.
4) Mohd Ismail Nazri and A.M. Zin, Emulation Network Analyzer Development for Campus Environmetn and Comparison between OPNET Application and Hardware Network Analyzer. European Journal of Scientific Research, 2008. 24(2): p. 270-291.
5) Mohd Ismail Nazri and A.M. Zin, Evaluation of Software Network Analyzer Prototyping Using Qualitative Approach. European Journal of Scientific Research, 2009. 26(3): p. 170-182.
6) Sood, A., Network Design by using OPNET™ IT GURU Academic Edition Software. Rivier Academic Journal, 2007. 3(1).
7) Sjögren, H.R.C., Voice over IP, simulated IP-network, in School of Mathematics and Systems Engineering. 2008, Växjö University.
8) Chang, W.K. and H. S, Evaluating the performance of a web site via queuing theory, in software quality-ECSQ. 2002, springer-Verlag, Berlin, Heideberg: Helsinki, Finland. p. 63-72.
9) Mohd Nazri Ismail and M.T. Ismail, Analyzing of Virtual Private Network over Open Source Application and Hardware Device Performance. European Journal of Scientific Research, 2009. 28(2): p. 215-226.
10) Nicolas Dubois and B. Fondeviole, Method for Control Management Based on a Routing Protocol. US Patent, 2008. No: US 2008/0025333 Al.
11) Richard Scott Curtis and J.D. Forrester, System, Method and Program for Network Routing. US Patent, 2008. No: US 2008/0317056 Al.
12) Thomas P. Chu, R.N. and Y.-T. Wang, Automatically Configuring Mesh Groups in Data Networks. US Patent, 2010. No: US 2010/0020726 Al.
13) Xiaode Xu, M.S. and D. Shah, Routing Protocol with Packet Network Attributes for Improved Route Selection. US Patent, 2009. No: US 2009/0059908 Al
14) Rosenberg, J., Peer-to-Peer Network including Routing Protocol Enhancement. US Patent, 2009. No.: US 2009/0122724 Al.
15) Bruce COLE and A.J. Li, Routing Protocols for Accommodating nodes with Redundant Routing Facilities. US Patent, 2009. No: US 2009/0219804 Al.
16) Russell I. White, S.E.M., James L. Ng, and Alvaro Enrique Retana, Determining an Optimal Route Advertisement in a Reactive Routing Environment. US Patent, 2009. No.: US 2009/0141651 Al.
17) Paul J. Fong, E.K., David Gray, et.al, Configuring Cisco Voice Over IP, ed. S. Edition.
18) C. Demichelis and P. Chimento, IP Packet Delay Variation Metric for IP Performance Metrics (IPPM). Request for Comments: 3393, 2002.
19) Pedrasa, J.R.I. and C.A.M. Festin, An Enhanced Framing Strategy for Jitter Management, in TENCON 2005 IEEE Region 10. 2005: Melbourne, Qld. p. 1-6.

An Empirical Study on Data Mining Applications

P.Sundari[1] Dr.K.Thangadurai[2]

GJCST Computing Classification
H.2.8, J.1

***Abstract*-The wide availability of huge amounts of data and the need for transforming such data into knowledge influences towards the attraction of IT industry in data mining. During the early years of the development of computer techniques for business, IT professionals were concerned with designing databases to store the data so that information could be easily and quickly accessed. The restrictions are storage space and the speed of retrieval of the data. Needless to say, the activity was restricted to a very few, highly qualified professionals. Then came an era when Database Management System simplified the task. Thus almost any business such as small, medium or large scale began using computers for day - to- day activities. Now what is the use of all this data? Up to the early 1990's the answer to this was "NOT much". No one was really interested in utilizing data, which was accumulated during the process of daily activities. As a result a new discipline in Computer Science, Data Mining gradually evolved. Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in financial transactions to discover illegal activities or analyzing genome sequences. This paper deals with the application of data mining in various fields in our day to day life.**

Keywords-Data Mining, Targeted Marketing, Market Based Analysis, Customer Relations

I. Introduction

Data Mining – An Overview

Data mining refers to extracting knowledge from large amounts of data. The data may be spatial data, multimedia data, time series data, text data and web data. Since Data mining is a young discipline with wide and diverse applications. In this paper we will discuss a few application domains of data mining such as Science and Engineering, Banking, Business, Telecommunication and Surveillance.

Data mining is the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data. Using information contained within data warehouse, data mining can often provide answers to questions about an organization that a decision maker has previously not thought to ask.

- ❖ Which products should be promoted to a particular customer? – Targeted Marketing
- ❖ What is the probability that a certain customer will leave for a competitor? – Customer Relationship Management
- ❖ What is the appropriate medical diagnosis for this patient? – Bio medical
- ❖ What is the likelihood that a certain customer will default or pay back a loan? – Banking
- ❖ Which products are bought most often together? – Market Basket Analysis
- ❖ How to identify fraudulent users in telecommunication industry? – Fraudulent pattern analysis

These types of questions can be answered quickly and easily if the information hidden among the huge amount of data in the databases can be located and utilized. We will discuss about the applications of data mining in the following paragraphs.

II. Applications of Data Mining

Although a large variety of data mining scenarios can be discussed, for the purpose of this paper the applications of data mining are divided into the following categories:

- ➢ Science and Engineering
- ➢ Business
- ➢ Banking
- ➢ Telecommunication
- ➢ Spatial data mining
- ➢ Surveillance

II. (A) Science and Engineering

The data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

i) Biomedical and DNA Data analysis

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the identification and study of human genome by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment. It is challenging to identify particular gene sequence patterns that play roles in various diseases. DNA data analysis is done in the following ways.[5]

About-[1] Department of Computer Science, Government Arts College (Women) , Krishnagiri- 635 001, India
(e-mail;sundaripalanisamy@yahoo.co.in)
About-[2] Department of Computer Science, Government Arts College.
(Men) , Krishnagiri- 635 001, India

- Semantic integration of heterogeneous, distributed genome databases
- Similarity search and comparison among DNA sequences
- Identification of co occurring gene sequences
- Path analysis includes linking genes to different stages of disease development
- Visualization tools and genetic data analysis
- The data mining technique that is used to perform this task is known as Multifactor Dimensionality Reduction.[3]

In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents.[7] Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.[8]

ii) Education

The other area of application for data mining in science/engineering is within educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning and to understand the factors influencing university student retention.[6] A similar example of the social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate Institutional memory.

iii) Electrical power engineering

In the area of electrical power engineering, data mining techniques have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on the insulation's health status of the equipment. Data clustering such as Self-Organizing Map (SOM) has been applied on the vibration monitoring and analysis of transformer On-Load Tap-Changers(OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for the exact same tap position. SOM has been applied to detect abnormal conditions and to estimate the nature of the abnormalities.[4]

Data mining techniques have also been applied for Dissolved Gas Analysis (DGA) on power transformers. DGA, as a diagnostics for power transformer, has been available for many years. Data mining techniques such as SOM has been applied to analyze data and to determine trends which are not obvious to the standard DGA ratio techniques such as Duval Triangle.[4]

Data mining technique is used to an integrated-circuit production line[2]. The data mining technique is applied in decision analysis to the problem of die-level functional test. Experiments demonstrate the ability of applying a system of mining historical die-test data to create a probabilistic model of patterns of die failure which are then utilized to decide in real time which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products.

b) Banking

Banking data mining applications may, for example, need to track client spending habits in order to detect unusual transactions that might be fraudulent. Most banks and financial institutions offer a wide variety of banking services (such as checking, saving, and business and individual customer transactions), credit (such as business, mortgage, and automobile loans), and investment services (such as mutual funds) [5]. It has also offer insurance services and stock services. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money. The following methods are used for financial data analysis.

- Loan payment prediction and customer credit policy analysis
- Classification and clustering of customers for targeted marketing
- Detection of money laundering and other financial crimes

c) Business

Retail industry collects huge amount of data on sales, customer shopping history, goods transportation and consumption and service records and so on. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail industry provides a rich source for data mining. Retail data mining can help identify customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business [5]. A few examples of data mining in the retail industry are as follows.

- Design and construction of data warehouses based on benefits of data mining
- Multidimensional analysis of sales, customers, products, time and region:

The multi feature data cube is a useful data structure in retail data analysis.

Another example of data mining, often called the market basket analysis, relates to its use in retail sales. If a clothing store records the purchases of customers, a data-mining system could identify those customers who favors silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical or inexact rules may also be present within a database. In a manufacturing application, an inexact rule may state that 73% of products which have a specific defect or problem will develop a secondary problem within the next six months.

Market basket analysis has also been used to identify the purchase patterns of the Alpha consumer. Alpha Consumers are people that play key roles in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society. Analyzing the data collected on these type of users has allowed companies to predict future buying trends and forecast supply demands.

Data Mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich history of customer transactions on millions of customers dating back several years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

- Analysis of the effectiveness of sales campaigns:
- Customer retention – analysis of customer loyalty

There are a wide variety of data mining applications available, particularly for business uses, such as Customer Relationship Management (CRM). Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can be used to investigate changes in customer consumption and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new customers. These applications enable marketing managers to understand the behaviors of their customers and also to predict the potential behavior of prospective customers. A data mining technique may assist the prediction of future customer retention. For example, a company may decide to increase prices, and could use data mining to predict how many customers might be lost for a particular percentage increase in product price.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.[1]

d) Telecommunication

The telecommunication industry offers local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service.

e) Spatial data mining

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

Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for applied GIS-based decision-making. Recently, the task of integrating these two technologies has become critical, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information hidden there. Among those organizations are:

Offices requiring analysis or dissemination of geo-referenced statistical data.

Public health services searching for explanations of disease clusters.

Environmental agencies assessing the impact of changing land-use patterns on climate change.

Geo-Marketing companies doing customer segmentation based on spatial location.

f) Surveillance

Data Mining is used by intelligence agencies like FBI and CIA to identify threats of terrorism. After the 9/11 incident it has become one of the prime means to uncover terrorist plots. However this led to concerns among the people as data collected for such works undermines the privacy of a large number of people.

Two plausible data mining techniques in the context of combating terrorism include "pattern mining" and "subject-based data mining".

i) Pattern mining

"Pattern mining" is a data mining technique that involves finding existing patterns in data. Pattern mining is a tool to

identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise."[9][10][11] Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search techniques.

ii) Subject-based data mining

"Subject-based data mining" is a data mining technique involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."[9]

g) Text Mining and Web Mining

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established.

An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e- commerce and many other information services. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites. For example the search engines work on the principle of data mining.

III. Need of Data Mining

The massive growth of data is due to the wide availability of data in automated form from various sources as WWW, Business, science, Society and many more. Data is useless, if it cannot deliver knowledge. That is why data mining is gaining wide acceptance in today‘s world. A lot has been done in this field and lot more need to be done.

IV. Conclusion

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. The aim of the paper is the study of application domains of Data Mining such as science and engineering, banking, business and telecommunication. Although data mining is a young field with many issues that still need to be researched in depth. The diversity of data, data mining tasks and approaches poses many challenging research issues in data mining. The design of data mining languages, the development of efficient and effective data mining methods, the construction of interactive and integrated data mining environments and the application of data mining techniques to solve large application problems are important tasks for data mining researchers.

V. References

1) Ellen Monk, Bret Wagner (2006). Concepts in Enterprise Resource Planning, Second Edition. Thomson Course Technology, Boston, MA. ISBN 0-619-21663-8. OCLC 224465825.
2) Tony Fountain, Thomas Dietterich & Bill Sudyka (2000) Mining IC Test Data to Optimize VLSI Testing, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (pp. 18-25). ACM Press.
3) Xingquan Zhu, Ian Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities. Hershey, New Your. pp. 18. ISBN 978-159904252-7.
4) a b A.J. McGrail, E. Gulski et al.. "Data Mining Techniques to Asses the Condition of High Voltage Electrical Plant". CIGRE WG 15.11 of Study Committee 15.
5) Jiawei Han & Micheline Kamber. (2001) Data Mining: Concepts and Techniques , Morgan Kaufmann publishers, CA,USA.
6) J.F. Superby, J-P. Vandamme, N. Meskens. "Determination of factors influencing the achievement of the first-year university students using data mining methods". Workshop on Educational Data Mining 2006.
7) Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, De Freitas RM. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998 Jun;54(4):315-21.
8) Norén GN, Bate A, Hopstadius J, Star K, Edwards IR. Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient Records. Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining SIGKDD 2008, pages 963-971. Las Vegas NV, 2008.

9) a b National Research Council, Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment, Washington, DC: National Academies Press, 2008.

10) R. Agrawal et al., Fast discovery of association rules, in Advances in knowledge discovery and data mining pp. 307-328, MIT Press, 1996.

11) Stephen Haag et al. (2006). Management Information Systems for the information age. Toronto: McGraw-Hill Ryerson. pp. 28. ISBN 0-07-095569-7. OCLC 63194770.

A Novel Decision Scheme for Vertical Handoff in 4G Wireless Networks

E. Arun[1] R.S Moni[2]

GJCST Computing Classification
C.2.1, F.4.2

***Abstract*-Future wireless networks will consist of multiple heterogeneous access technologies such as UMTS, WLAN, and Wi-Max. These technologies differ greatly regarding network capacity, data rates, and other various parameters such as power consumption, Received Signal Strength, and coverage areas. This paper presents two Handoff Decision schemes for heterogeneous networks. A good handoff decision could avoid the redundant handoffs and reduce the packet lose. First scheme makes use of a score function to find the best network at best time from a set of neighboring networks. Score function uses bandwidth, Received Signal Strength (RSS) and access fee as its parameters. Second scheme makes use of classic triangle problem to find the best network from a set of neighboring networks. This problem considers three parameters bandwidth, Received Signal Strength (RSS) and access fee as the three sides of a triangle. If an equilateral triangle is obtained with these parameters of a network then that network will be the best among the set of networks. The best decision model meets the individual user needs but also improve the whole system performance by reducing the unnecessary handoffs.**

Keywords-MIHF, Received Signal Strength, Mobility Management, vertical handoff ,

I. Introduction

Currently, there are various wireless networks deployed around the world. Examples include second and third generation (3G) of cellular networks (e.g., GSM/GPRS, UMTS, CDMA2000), wireless local area networks WLANs (e.g., IEEE 802.11a/b/g), and personal area networks (e.g., Bluetooth). All these wireless networks are heterogeneous in sense of the different radio access technologies. From this fact, it follows that no access technology or service provider can offer ubiquitous coverage expected by users requiring connectivity anytime and anywhere. The actual trend is to integrate complementary wireless technologies with overlapping coverage, to provide the expected ubiquitous coverage and to achieve the ―Always Best Connected" (ABC) concept The ABC concept allows the user to use the best available access network. In order to accomplish the integration and inter-working between heterogeneous wireless networks and the ABC concept, many challenging research problems have to be solved, taking into account that all these new wireless technologies were designed without considering any interworking among them In heterogeneous wireless networks, mobile devices or mobile stations will be equipped with multiple network interfaces to access different wireless networks. Users will expect to continue their connections without any disruption when they move from one network to another. This important process in wireless networks is referred to as handoff or handover.

Handoff process among networks using different access technologies is defined as vertical handoff (VHO) [1]. Such a process of changing the connections among different types of wireless and mobile networks is called the vertical handoff. Obviously, the network selection and the vertical handoff decision are two important processes in an integrated wireless and mobile network. Handoff process is initiated by change in different factors like Received Signal Strength (RSS), Signal to Noise Ratio (SNR) etc. When these factors fall bellow the threshold value the Mobile Node (MN) has to search for another AP having RSS greater than threshold value [2, 3]. Wang et al. introduce the policy enabled handoff in [4], which was followed by several papers on similar approaches. Policy enabled handoff systems separates the decision making (i.e. which is the ―best" network and when to handoff) from the handoff mechanism. Smart Decision Model [5] smartly performs vertical handoff among available network interfaces. Using a well-defined score function, the proposed model can properly handoff to the ―best" network interface at the ―best" moment according to the properties of available network interfaces, system configurations / information, and user preferences. A handoff decision scheme with guaranteed QoS [6] for heterogeneous networks make the decision according to the user‗s communicating types and the performance of the networks. A generic vertical handoff decision function [7] proposed considering the different factors and metric qualities that give an indication of whether or not a handoff is needed. The decision function enables devices to assign weights to different network factors such as monetary cost, quality of service, power requirements, personal preferences etc. A decision strategy [8] considers the performance of the whole system while taking VHO decisions by meeting individual needs. This decision strategy select the best network based on the highest received signal strength (RSS) and lowest Variation of Received Signal Strength (VRSS). Thus it ensures the high system performance by reducing the unnecessary handoffs. Nasser et al. [9] proposed a VHO decision (VHD) method that simply estimates the service quality for

About-[1]Assistant Professor, Dept of Computer Science & Engineering, Noorul Islam University, Thuckalay, Tamil Nadu, India (e-mail;arunsedly@yahoo.com)
About-[2]Senior Professor, Dept of Electronics &Communication Engineering, Noorul Islam University, Thuckalay, Tamil Nadu, India (e-mail;moni2006_r.s@yahoo.co.in)

available networks and selects the network with the best quality. However, there still lie ahead many challenges in integrating cellular networks and WLANs.

This paper is organized as follows. In Section II, we introduce our proposed system model for an integrated wireless and mobile network. In Section III, different handoff decision strategies are presented. In Section IV, we analyze the performance of the proposed strategy. Finally, we conclude this paper in Section V.

II. System Model

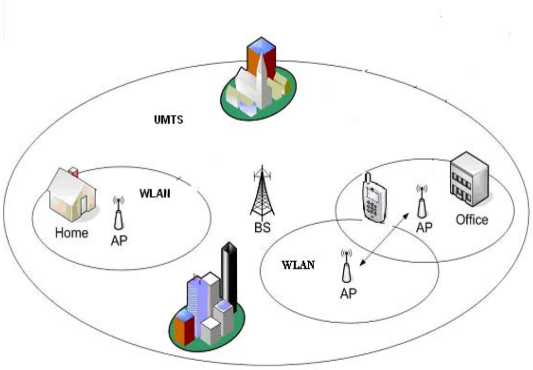


Fig 1 Vertical handoff in heterogeneous networks

As shown in the above figure an MN can be existing at a given time in the coverage area of an UMTS alone. However, due to mobility, it can move into the regions covered by more than one access network, i.e., simultaneously within the coverage areas of, for example, an UMTS BS and an IEEE 802.11 AP. Multiple IEEE 802.11 WLAN coverage areas are usually contained within an UMTS coverage area. A Worldwide Interoperability for Microwave Access (WiMAX) coverage area can overlap with WLAN and/or UMTS coverage areas. In dense urban areas, even the coverage areas of multiple UMTS BSs can overlap. Thus, at any given time, the choice of an appropriate attachment point (BS or AP) for each MN needs to be made. These access technologies have different bandwidth, power consumption, RSS threshold, data rate, jitter, delay etc. So during handoff it is required to find the best network according to user preferences. At the hotspots APs are made available. When the Received Signal Strength of an AP goes low below some threshold value the Mobile Host has to find another best network considering bandwidth, RSS, access fee as parameters. Each of these parameters is given a weight according to preferences. If any of the best AP s are not available handoff has to be performed to Base Station of UMTS. Thus, multiple access technologies and multiple operators are typically involved in Network Selection Decision. The Network Selection decision making algorithm is implemented in Network selection decision Controllers located in access networks. Decision input for NSDCs will be obtainable via the MIHF. The MIHF of NSDC facilitates standard based message exchanges between various access networks or attachment points to share information about the traffic load, bandwidth available, RSS and other network capabilities of each AP. NSDC obtains LLT s from MN via MIHF. LLT regarding MN indicates two possibilities a) RSS for an MN dropped below some specific threshold while MN in service at an AP b) RSS for one or more APs exceeded to a specific threshold while MN in service at BS. Usually AP is preferred attachment point than BS since AP is associated with higher bandwidth cost and higher data rate. When NSDC obtains LLT it executes Network selection decision algorithm and find the best AP, if no other best APs are found for handoff select cellular network as the best available network.

III. NETwork selection Decision Making Algorithms

Most existing network selection strategies only focused on the individual user‘s needs. Motivation of this paper is to design a network-selection strategy from a system‘s perspective, and the network-selection strategy can also meet a certain individual user‘s needs. In the following, we discuss how our proposed network-selection strategy works.

A. Algorithm

1) *Handoff Initiation:*

MN can be in service with AP or BS. When the RSS strength goes low below some threshold value or when the RSS strength in any of the AP goes above some threshold value when the MN is in service with BS, the MN has to find a best network to which it has to perform handoff .When RSS goes low MN gives Link layer trigger to Network Selection Decision Controller in the network in which the MN currently connects to. Thus the handoff process is initiated.

2) Handoff Decision:

When handoff process is initiated, the Network Selection decision controller collects the condition of each neighboring network via Media Independent Handover Function (MIHF) and executes Network Selection Decision Controller (NSDC) algorithm. The algorithm first calculates the score of the current network and compares the score with each of the neighboring network‗s score. The score of the neighboring networks is calculated only if all the parameters have satisfying value to accept a Mobile Host. Our proposed network-selection strategy prefers a call to be accepted by a network with lower traffic load and stronger received signal strength, which can achieve better traffic balance among different types of networks and good service quality. Consequently, we define a score function to combine these two factors-the traffic load and the received signal strength. Therefore, the score to use a network Ni for a call is defined as the score function used is the following:

$$Score = \sum_{j=1}^{k} W_j Norm_j \qquad (1)$$

k is the number of parameters. Wj is the weight assigned to the parameter j. Normj is the normalized value of the parameter j. If any of the network with higher score is available handoff to that particular network or if any of the network with optimum score is not available handoff to BS.

$$Score_i = wg.G_i + ws.S_i + wf.F_i \qquad (2)$$

where Gi is the complementary of the normalized utilization of network Ni, Ri is the relative received signal strength from network Ni, Fi is the normalized access fee of network Ni and wg (0 ≤ wg ≤ 1) ws (0 ≤ ws ≤ 1),wf (0 ≤ wf ≤ 1), are the weights that provide preferences to Gi, Si, Fi respectively. The larger the weight of a specific factor, the more important that factor is to the user and vice versa The constraint between wg ,ws and wf is given by

$$wg + ws+wf = 1 \qquad (3)$$

Even though we could add the different factors in the VHDF to obtain network score, each network parameter has a different unit, which leads to the necessity of normalization. The complementary of normalized utilization Gi is defined by

$$Gi = \frac{B_{if}}{B_i} \qquad (4)$$

where Bif is the number of available bandwidth units in network Ni, Bi is the total number of bandwidth units in network Ni.

In general, stronger received signal strength indicates better signal quality. Therefore, an originating call prefers to be accepted by a network that has higher received signal strength. However, it is difficult to compare the received signal strength among different types of wireless and mobile networks because they have different maximum transmission power and receiver thresholds. As a result, we propose to use relative received signal strength to compare different types of wireless and mobile networks. Si in (2) is defined by

$$Si = \frac{P_i^c - P_i^{th}}{P_i^{\max} - P_i^{th}} \qquad (5)$$

where P_i^c is the current received signal strength from network Ni, P_i^{th} is the receiver threshold in network Ni, and $P_i^{\max}$ is the maximum transmitted signal strength in network Ni. It is to note that we only consider the path loss in the radio propagation model. Consequently, the received signal strength (in decibels) in network Ni is given by

$$P_i^c = P_i^{\max} - 10\gamma \log(r_i) \qquad (6)$$

where ri is the distance between the mobile user and the BS (or AP) of network Ni, and γ is the fading factor . Therefore, the receiver threshold in network Ni is given by

$$P_i^{th} = P_i^{\max} - 10\gamma \log(R_i) \qquad (7)$$

The relative received signal strength from network Ni is rewritten as

$$S_i = 1 - \frac{\log(r_i)}{\log(R_i)} \qquad (8)$$

Ri is the radius of cell of network i

Access fee Φi is given by

$$\Phi_i = \frac{(1-\varphi_i)}{\varphi_{\max}} \qquad (9)$$

where *φ*max is the highest access fee that the mobile user likes to pay, and *φi* is the access fee to use network *Ni*. The mobile user does not connect to a network that charges more than *φ*max. If an originating call has more than one connection option, the score of all candidate networks are calculated by using the score function in (2). The originating call is accepted by a network that has the largest score, which indicates the –best” network. If there is more than one –best” network, the originating call is randomly accepted by any one of these –best” networks.

Flow chart

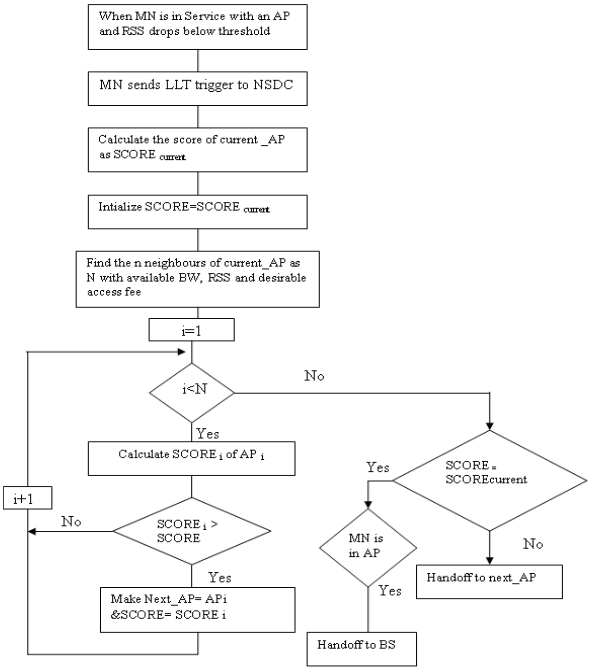

When MN is in Service with an AP and RSS drops below threshold
MN sends LLT trigger to NSDC
Calculate the score of current _AP as SCORE current
Intialize SCORE=SCORE current
Find the n neighbours of current_AP as N with available BW, RSS and desirable access fee
i=1
i<N
No
Yes
Calculate SCORE i of AP i
i+1
No
SCORE i > SCORE
Yes
Make Next_AP= APi &SCORE= SCORE i
Yes
SCORE = SCOREcurrent
MN is in AP
Yes
Handoff to BS
No
Handoff to next_AP

Fig 2: Handoff decision Algorithm 1

Here this algorithm checks only if bandwidth is available and not checking it greater than threshold. As the available bandwidth decreases i.e. the load increases there is more chance for the RSS to go low. Thus the call dropping probability increases and holding time decreases. In this algorithm if any of the parameters have greater value the score increases even if others have less value.

B. Algorithm 2

1) Handoff Initiation

MN can be in service with AP or BS. When the RSS strength goes low below some threshold value or when the RSS strength in any of the AP goes above some threshold value when the MN is in service with BS, the MN has to find a best network to which it has to perform handoff .When RSS goes low MN gives Link layer trigger to Network Selection decision controller in the network in which the MN currently connects to. Thus the handoff process is initiated.

2) Handoff Execution:

Handoff execution is based on classic triangular problem. According to triangular problem we consider triangles representing the conditions of networks. Each side of the triangle corresponds to each parameter. The parameters this problem considers in this paper are Received Signal Strength, Bandwidth and Access cost. If all the parameters have desired value (value MN expects) then the resultant triangle will be equilateral (S1=S2=S3=a, three sides equal) and if two of the parameters have desired value the triangle will be isosceles (S1≠S2=S3 or S1=S2≠S3, two sides equal). If S1≠S2≠S3 then the triangle is scalene. The networks that give equilateral triangle and isosceles will be in candidate list 1 and candidate list 2 respectively. Select one network from list1 as best network and if list1 is empty select best network from list2. Then perform handoff to the selected best network. If both lists are empty handoff to BS.

Flow chart

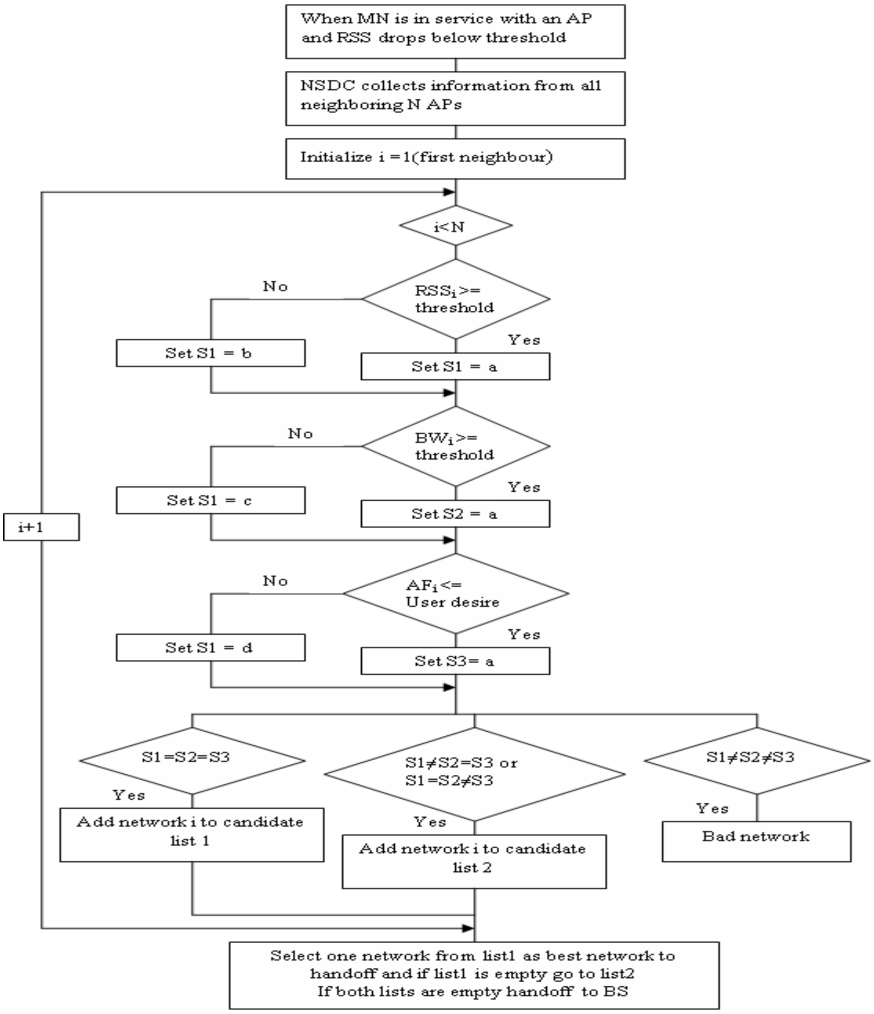

When MN is in service with an AP and RSS drops below threshold
NSDC collects information from all neighboring N APs
Initialize i =1(first neighbour)
i<N
No
RSSi>= threshold
Yes
Set S1 = b
Set S1 = a
No
BWi>= threshold
Yes
Set S1 = c
Set S2 = a
i+1
No
AFi<= User desire
Yes
Set S1 = d
Set S3= a
S1=S2=S3
Yes
Add network i to candidate list 1
S1≠S2=S3 or S1=S2≠S3
Yes
Add network i to candidate list 2
S1≠S2≠S3
Yes
Bad network
Select one network from list1 as best network to handoff and if list1 is empty go to list2 If both lists are empty handoff to BS

Fig 3: Handoff decision Algorithm 2

RSS can be measured as

$$P_i^c = P_i^{\max} - 10\gamma \log(r_i) \qquad (10)$$

where P_i^c is the current received signal strength from network *Ni*, *ri* is the distance between the mobile user and the BS (or AP) of network. $P_i^{\max}$ is the maximum transmitted signal strength in network Ni γ is the fading factor

Bandwidth is given by
Available Bandwidth of the network = Bandwidth of the network − sum of Bandwidth used by all MNs Attached to the network.

Access Fee is the fee that is assigned to each network usage. It may vary from network to network. User usually prefers the low network fee.

IV. PERFORMANCE ANALYSIS

Simulations have been performed for the 3G cell overlay structure. In this scenario three networks of different data rates co-exist in the same wireless service area. Network 1 and Network 2 represent 802.11b wireless LANs, with bandwidths of 2Mbps and 1Mbps, respectively. Network 3 is modeled as a UMTS network, which supports multiple users simultaneously.

The expected graphs are shown below

Bandwidth	10	20	30	40	50	60	70	80	90	100
Holding Time algorithm 1	2.5	4.5	5.7	6.1	6.3	6.5	6.9	7	7	7
Holding Time algorithm 2	3.5	5.5	6.5	7	7.5	8	8.5	9	9.5	9.5

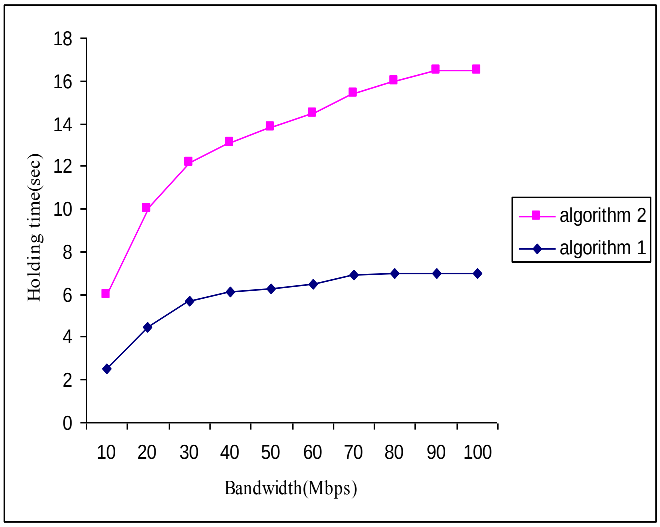


Fig 4: Holding time Vs RSS

RSS	5	10	20	30	40	50	60
algorithm 2 call dropping probability	0.5	0.4	0.3	0.25	0.2	0.1	0.09
algorithm 1 call dropping probability	0.8	0.75	0.7	0.65	0.55	0.4	0.3

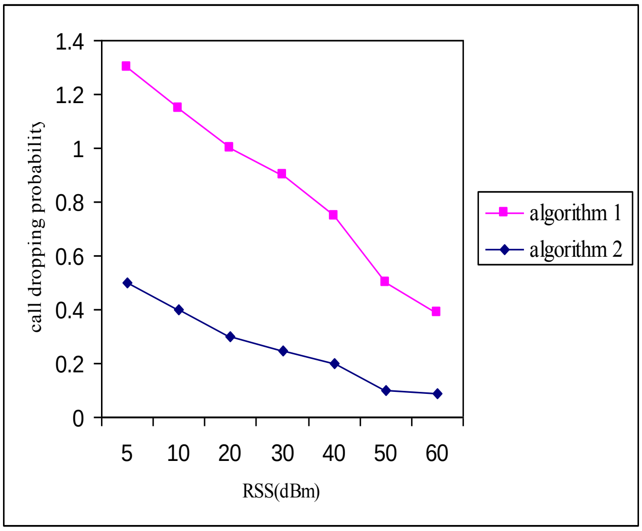


Fig 6: Call dropping probability Vs RSS

V. CONCLUSION

Thus this paper describes two different handoff decision algorithms. First algorithm uses a score function to find the best network at best time from a set of neighboring networks. Second algorithm uses classic triangle problem to find the best network from a set of neighboring networks. If an equilateral triangle is obtained with three parameters of a network then that network will be the best among the set of networks. Since the second algorithm performs handoff only if the constraints are above the threshold value. The call dropping probability is reduced and holding time is increased.

VI. REFERENCES

1) Enrique Stevens-Navarro, Ulises Pineda-Rico, and Jesus Acosta-Elias ―Vertical Handover in beyond Third Generation (B3G) Wireless Networks" International Journal of Future Generation Communication and Networking, pp. 51-58, 2008
2) K.Ayyappan and P.Dananjayan ―RS measurement for vertical handoff in heterogeneous network", Journal of Theoretical and Applied Information Technology, pp. 989-994 , 2005
3) Kemeng Yang, Iqbal Gondal, Bin Qiu and Laurence S. Dooley ―Combined SINR Based Vertical Handoff Algorithm for Next Generation Heterogeneous Wireless Networks" Global Telecommunications Conference, 2007. GLOBECOM '07, pp. 4483 – 4487, Nov 2007, Digital Object Identifier 10.1109/GLOCOM.2007.852

4) Wang, R. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks", WMCSA 99, Feb1999, pp. 51-60, Digital Object Identifier 10.1109/MCSA.1999.749277

5) L-J. Chen, T. Sun, B. Chen, V. Rajendran, and M. Gerla. "A Smart Decision Model for Vertical Handoff." The 4th ANWIRE International Workshop on Wireless Internet and Reconfigurability (ANWIRE 2004). May 2004.

6) Ying-Hong Wang, Chih-Peng Hsu, Kuo-Feng Huang, Wei-Chia Huang"Handoff decision scheme with guaranteed QoS in heterogeneous network" pp 138-143,2008Digital Object Identifier 10.1109/UMEDIA.2008.4570879

7) Ahmed Hasswa, Nidal Nasser, Hossam Hassanein ―Generic Vertical Handoff Decision Function for Heterogeneous Wireless Networks" Wireless and Optical Communications Networks, 2005. WOCN 2005, pp 239-243,Mar 2005 , Digital Object Identifier 10.1109/WOCN.2005.1436026

8) Shen,W.;Zeng,Q.-A. ―ANovel Decision Strategy of Vertical Handoff in Overlay Wireless Networks" Fifth IEEE International Symposium on Network Computing and Applications, 2006 ,pp 227-230 Digital Object Identifier 10.1109/NCA.2006.5

9) Summary: In an overlay wireless network, a mobile user can connect to different radio access networks if it is equipped with appropriate network interfaces. When the mobile user changes its connection between different radio access networks, a vertical handof....N. Nasser, A. Hasswa, and H. Hassanein, ―Handoffs in fourth generation heterogeneous networks," IEEE Commun. Mag., vol. 44, no. 10, pp. 96–103, Oct. 2006, Digital Object Identifier 10.1109/MCOM.2006.1710420

10) Olga Ornond, Philip Perry and John Murphy"Network Selection Decision in Wireless Heterogeneous Networks" 2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications Volume 4,pp 2680 – 2684 Sept 2005, Digital Object Identifier 10.1109/PIMRC.2005.1651930

11) Wei Shen, and Qing-An Zeng, ―Cost-Function-Based Network Selection Strategy in Integrated Wireless and Mobile Networks," IEEE Trans. Veh. Technol., vol. 57, no. 6, pp. 3778–3788, Nov. 2008. Digital Object Identifier 10.1109/TVT.2008.917257.

Hybrid Approach for Template Protection in Face Recognition System

Sheetal Chaudhary[1] Rajender Nath[2]

GJCST Computing Classification
I.2.10, I.2.7, I.4.6

***Abstract*-Biometrics deals with identifying individuals with the help of their biological (physiological and behavioral) data. The security of biometric systems has however been questioned and previous studies have shown that they can be fooled with artificial artifacts. Also biometric recognition systems face challenges arising from intra-class variations and attacks upon template databases. To tackle such problems, a hybrid approach for liveness detection and protecting templates in face recognition system is proposed. Here, the system captures input face image in three different poses (left, front, right) based upon the order chosen by the random select module. This approach will perform live face detection based upon complete body movement of the person to be recognized and template protection by randomly shuffling and adding the components of feature set resulting after fusion of three poses of input face image. It overcomes the limitations imposed by intra-class variations and spoof attacks in face recognition system. The resulting hybrid template will be more secure as original biometric template will not be stored in the database rather it will be stored after applying some changes (shuffling and addition) in its components. Thus the proposed approach has higher security and better recognition performance as compared to the case when no measures are used for live face check and template protection in database.**

Keywords-Liveness detection, template protection, face recognition, multiple sample fusion, eigen-coefficients

I. Introduction

The term biometrics is derived from the Greek words bios and metron which translates as life measurement. Biometrics are not secrets and therefore should be properly protected. A good biometrics system should depend not only on security of biometric data but the authentication process must also check for liveness of the biometric data. People leave fingerprints behind on everything they touch, and the iris can be observed anywhere they look. Our facial images are recorded every time we enter a bank, railway station, and supermarket [1]. Once biometric measurements are disclosed, they cannot be changed (unless the user is willing to have an organ transplant). The only way how to make a system secure is to make sure that the data presented came from a real person and is obtained at the time of authentication. Liveness detection in a biometric system means the capability for the system to detect, during enrollment and identification/verification, whether or not the biometric sample presented to the system is alive or not. It must also check that the presented biometric sample belongs to the live human being who was originally enrolled in the system and not just any live human being. It is well known that fingerprint systems can be fooled with artificial fingerprints, static facial images can be used to fool face recognition systems, and static iris images can be used to fool iris recognition systems [2].

Multimodal biometric systems consolidate the evidence presented by multiple biometric sources of information and are expected to be more reliable due to the presence of multiple, fairly independent pieces of evidence [3]. Intra-class variations in face recognition system can be overcome with multimodal biometric systems. Figure 1 is showing intra-class variation associated with an individual‘s face image. Due to change in pose, face recognition system will not be able to match these 3 images successfully, even though they belong to the same individual [4]. A Multibiometric system can be classified into five categories (multi-sensor, multi-algorithm, multi-instance, multi-sample and multimodal) depending upon the evidence presented by multiple sources of biometric information. Multi-sample system can be used to tackle intra-class variations. Here, a single sensor is used to acquire multiple samples of the same biometric trait in order to account for the variations that can occur in the trait. It is an inexpensive way of improving system performance since this system requires neither multiple sensors nor multiple feature extraction and matching modules [5] [6].

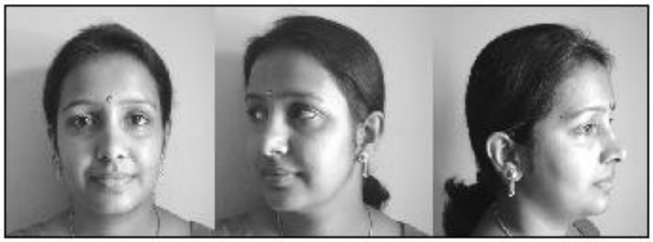


Fig.1: Intra-class variation associated with an individual‘s face image

One of the properties that make biometrics so attractive for authentication purposes is their invariance over time. One of the most vulnerabilities of biometrics is that once a biometric image or template is stolen, it is stolen forever and cannot be reissued, updated or destroyed [7]. One of the most potentially damaging attacks on a biometric system is

About-[1]University Research Scholar, Department Of Comp. Sc. & App. K.U., Kurukshetra, Haryana, India (e-mail;Sheetalkuk@Rediffmail.Com)
About-[2]Associate Professor, Department Of Comp. Sc. & App. K.U., Kurukshetra, Haryana, India (e-mail;rnath_2k3@rediffmail.com)

against the biometric template database. Attacks on the template can lead to the following three vulnerabilities: (i)
A template can be replaced by an impostor‘s template to gain unauthorized access, (ii) A physical spoof can be created from the template to gain unauthorized access to the system (as well as other systems which use the same biometric trait) and (iii) The stolen template can be replayed to the matcher to gain unauthorized access [8].

The proposed hybrid approach provides three main advantages: handles intra-class variation, performs live face check and provides protection against attacks on template database. The rest of the paper is organized as follows. Section 2 addresses the literature study. In section 3 face feature set extraction using PCA is discussed. In section 4 architecture of the proposed approach is presented. Section 5 discusses the advantage of proposed approach. Finally, the summary and conclusions are given in last section.

II. RELATED WORK

In recent years face recognition has received substantial attention from both research communities and the market, but still remained very challenging in real applications. A lot of face recognition algorithms have been developed during the past decades. Face recognition consists in localizing the most characteristic face components (eyes, nose, mouth, etc.) within images that depict human faces This step is essential for the initialization of many face processing techniques like face tracking, facial expression recognition or face recognition. Among these, face recognition is a lively research area where a great effort has been made in the last years to design and compare different techniques [9]. Hong and Jain [10] designed a decision fusion scheme to combine faces and fingerprint for personal identification. Brunelli and Falavigna [11] presented a person identification system by combining outputs from classifiers based on audio and visual cues. Face recognition algorithms are categorized into appearance based and model-based schemes. For appearance-based methods, three linear subspace analysis schemes are presented (PCA, LDA, and ICA) [12].The model-based approaches include Elastic Bunch Graph matching [13], Active Appearance Model [14] and 3D Morphable Model [15] methods. Among face recognition algorithms, appearance-based approaches are the most popular. These approaches utilize the pixel intensity or intensity-derived features.

The template protection schemes proposed in the literature can be broadly classified into two categories, feature transformation approach and biometric cryptosystem approach [8]. In the feature transformation approach, a transformation function is applied to the biometric template and only the transformed template is stored in the database. The same transformation function is applied to query features and the transformed query is directly matched against the transformed template. Depending on the characteristics of the transformation function, the feature transform schemes can be further categorized as salting and non-invertible transforms. In a biometric cryptosystem, some public information about the biometric template is stored. This public information is referred to as helper data and hence, biometric cryptosystems are also known as helper data-based methods. While the helper data does not reveal any significant information about the original biometric template, it is needed during matching to extract a cryptographic key from the query biometric features. Matching is performed indirectly by verifying the validity of the extracted key. Biometric cryptosystems can be further classified as key binding and key generation systems depending on how the helper data is obtained [16].

Liveness detection can be performed either at the acquisition stage, or at the processing stage. There are two approaches in determining if a biometric trait is alive or not; liveness detection and non-liveness detection [2]. Liveness detection, which aims at recognition of human physiological activities as the liveness indicator to prevent spoofing attack, is becoming a very active topic in field of fingerprint recognition and iris recognition, but efforts on live face detection are still very limited though live face detection is highly desirable. The most common faking way is to use a facial photograph of a valid user to spoof face recognition systems. Most of the current face recognition works with excellent performance are based on intensity images and equipped with a generic camera. Thus, an anti-spoofing method without additional device will be preferable, since it could be easily integrated into the existing face recognition systems [17] [18].

III. FEATURE EXTRACTION

Facial recognition is the identification of humans by the unique characteristics of their faces. It has attracted a lot of attention because of its potential applications. Among face recognition algorithms, appearance-based approaches (PCA, LDA, and ICA) are the most popular. These approaches utilize the pixel intensity or intensity-derived features [12].

In this paper, the PCA method using eigenfaces was adopted for face recognition. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. The main idea of the principal component analysis (or Karhunen-Loeve transform) is to find the vectors which best account for the distribution of face images within the entire image space. These vectors define the subspace of face images, which we call "*face space*". Because these vectors are the eigenvectors of the covariance matrix corresponding to the original face images, and because they are face like in appearance, we refer to them as ―*eigenfaces*". Eigenfaces are a set of eigenvectors used in the computer vision problem of human face recognition. The eigenfaces are the principal components of a distribution of faces, or equivalently, the eigenvectors of the covariance matrix of the set of face images, where each image with NxN pixels is considered a point (or vector) in N^2-dimensional space [19]. The idea of using principal components to represent human faces was developed by Sirovich and Kirby [20] and used by Turk and Pentland [21] for face detection and recognition. Eigenfaces are mostly used to:

(a) Extract the relevant facial information, which may or may not be directly related to face features such as the eyes,

nose, and lips. One way to do so is to capture the statistical variation between face images.
(b) Represent face images efficiently. To reduce the computation and space complexity, each face image can be represented using a small number of dimensions.
Mathematically, it is simply finding the principal components of the distribution of faces, or the eigenvectors of the covariance matrix of the set of face images, treating an image as a point or a vector in a very high dimensional space. Each eigenvector is accounting for a different amount of the variations among the face images. These eigenvectors can be imagined as a set of features that together characterize the variation between face images [19]

IV. PROPOSED APPROACH

Figure 2 shows the block diagram of the proposed approach for template protection in face recognition system. The main idea behind the proposed approach is to generate secure hybrid templates by integrating three different views (left, front and right) of input face image and then changing the components of resulting face feature set.

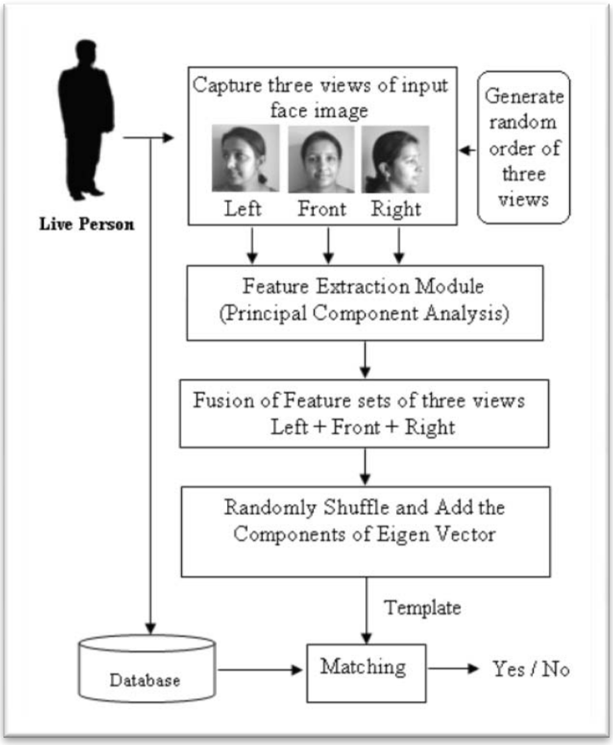

Live Person
Capture three views of input face image
Left
Front
Right
Generate random order of three views
Feature Extraction Module (Principal Component Analysis)
Fusion of Feature sets of three views Left + Front + Right
Randomly Shuffle and Add the Components of Eigen Vector
Template
Database
Matching
Yes / No

Fig. 2: Architecture of proposed approach for template protection in face recognition system

The proposed approach can be roughly divided into the following four steps:
A. Random selection of three Facial views (Left, Front, Right) to perform Liveness Detection.
B. Extraction of Feature sets of three Facial views
C. Fusion of Feature sets of three Facial views (Left + Front + Right).
D. Random shuffling and addition of components of Eigen vector (resulting after fusion).

A. Random selection of three Facial views

This step is responsible for performing liveness detection. Here, the person to be recognized is required to stand in front of camera which is focused upon full height of the person. Based upon the random order (LFR or LRF or FLR or FRL or RFL or RLF, L: Left; F: Front; R: Right) generated by the random select module as shown in fig. 2, the person is asked to move left or right or look at front. The camera is focused upon the entire body to examine the actual body movement but it will capture only the images of face in the order selected by the module. The module which generates random order of three views will detect whether the person is live or not by instructing the person to move left or right or look at front. Complete body movement is examined through camera and face images will be captured only if the person is live. To perform liveness detection, the random select module can be equipped with the following decision process which first checks liveness and then performs person recognition

```
if data = live
perform acquisition and extraction
else if data = not live
do not perform acquisition and extraction
```

B. Extraction of Feature sets of three Facial views

performs feature set extraction of three views (Left, Front, Right) of input face image by using PCA (appearance based) face recognition technique. PCA method is applied individually on each view of the face image to extract the corresponding feature set. When using PCA, each face image is assumed to be a 2-dimensional array of intensity values. It is represented as 1-dimensional vector by concatenating each row (or column) into a long thin vector. By projecting the face vector to the basis vectors, the projection coefficients are used as the feature representation of each face image. The PCA method using eigenfaces consists of the following two stages [10]

1) Training stage, in which a set of N face images are collected; eigenfaces that correspond to the M highest eigenvalues are computed from the data set; and each face is represented as a point in the M dimensional eigenspace, and
2) Operational stage, in which each test image is first projected onto the M-dimensional eigenspace; the M dimensional face representation is then deemed as a feature vector and fed to a classifier to establish the identity of the individual.

For each face image, we obtain a feature vector by projecting image onto the subspace generated by the principal directions of the covariance matrix. After applying the projection, the input vector (face) in an n-dimensional space is reduced to a feature vector in an m-dimensional subspace (M<< N) [9].
Thus, the feature vectors of three individual face views can be represented in terms of eigen vectors as described below
eigen vector for left face view $V_L = [a_1, a_2, a_3, a_4 \ldots a_m]$
eigen vector for front face view $V_F = [b_{1,} b_2, b_3, b_4 \ldots b_m]$

eigen vector for right face view $V_R = [c_{1,} c_2, c_{3,} c_4 \ldots c_m]$ where $V_{L,}$ $V_{F,}$ V_R represent the feature sets in terms of eigen-coefficients of three views of face image respectively.

C. *Fusion of Feature sets of three Facial views*

Fusion involves consolidating the evidence presented by two or more biometric feature sets of the same individual. This step performs fusion of feature sets of three face views of the same image at feature level [6]. Here, the three feature sets originate from the same feature extraction algorithm (PCA). Fusion of three face views is performed by just averaging them as given below

$$X = (VL + VF + VR)/3 \qquad (1)$$

The resulting fused eigen vector can be represented as X = [x1, x2, x3, x4…xm].

D. *Random shuffling and addition of components of Eigen vector*

This step in the proposed approach is responsible for performing changes in the eigen vector that is obtained after fusion of three feature sets. It will make the resulting template more secure. This step is illustrated in fig.3 below

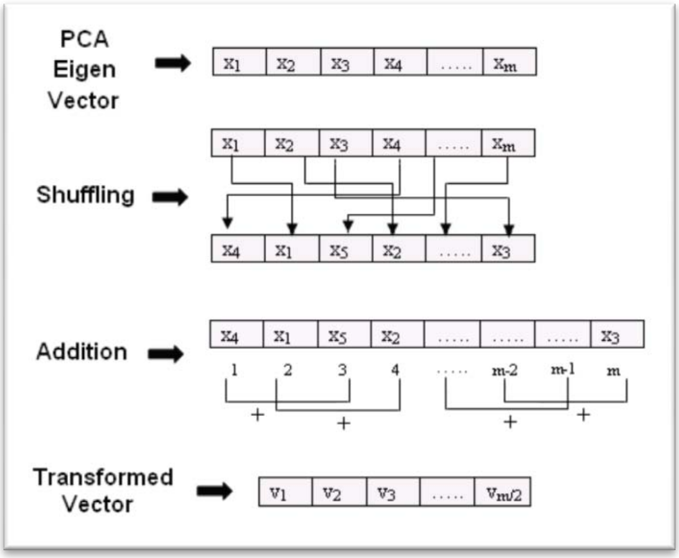


Fig. 3: Steps to generate secure hybrid template from input face images

The coefficients of eigen vector X are randomly shuffled. By shuffling, randomly chosen columns are interchanged and every time we can generate a new eigen vector.

$$X' = \text{Hshuffle}(X) \qquad (2)$$

where Hshuffle is the function which performs shuffling on the eigen vector X and X′ is the shuffled eigen vector. The number of coefficients in both X and X′ are same, shuffling just changes the order of columns. After that, addition among coefficients of shuffled eigen vector is performed in some order. Addition function is described below:

$$\text{Addition} = \sum_{p=1}^{p=m-2} [x_p + x_{p+2}], \quad (3)$$

after every two iteration p is incremented with 3.

Random shuffling of coefficients in the eigen vector that is obtained after fusion of three feature sets and addition among coefficients of shuffled eigen vector will generate the hybrid template that would be finally stored in the system database. The resulting hybrid template will contain half the no. of coefficients in the original eigen vector that was obtained in the previous step. The no. of coefficients are reduced by addition function. This approach will make the template more secure against spoof attacks and will take less memory in the database.

V. ADVANTAGE OF USING PROPOSED APPROACH

basic idea of the proposed approach is that instead of storing the original template in database, it is stored after performing fusion, shuffling and addition in the coefficients of eigen vector. The proposed approach offers advantage in terms of liveness detection, intra class variation, and template security by providing the ability to discard the stolen template information. Here, multiple samples of the same biometric trait (face) are captured in order to account for the intra class variations that can occur in the trait and for checking liveness of acquired biometric sample. This approach for liveness detection is natural, non-intrusive and no extra hardware is required. But it requires user collaboration by instructing the user to move left, right or stand in front of camera.

It provides template security by performing fusion of feature sets of three facial views (Left, Front, Right), random shuffling of eigen-coefficients in the fused eigen vector and addition among the shuffled eigen-coefficients. If the template is found to be compromised, the proposed approach provides the ability to discard it and reissue with new shuffling rules. In this way, with shuffling a number of eigen vectors can be generated. Also it is impossible for the attacker to convert the stolen template into the original face data (PCA eigen vector). It is well known that each eigenface represents certain characteristic features of faces and any original image can be reconstructed by combining the eigenfaces in right proportion. Hence, the original eigen vector is not stored in the database rather it is stored after applying shuffling rules and then adding the shuffled coefficients according to the addition function as discussed in the previous section. Addition reduces the size of the eigen vector by half and hence the final hybrid template generated will be compact and more secure. Thus the proposed scheme provides higher template security and better recognition performance as compared to the case when no measures for liveness detection and template protection are taken as in existing face recognition system using eigenfaces approach.

VI. CONCLUSION

Biometric template protection has become one of the important issues in deploying a practical biometric system. In this paper, a hybrid approach for template protection in face recognition system is proposed. This approach is based on the fusion of three different views (left, front, right view captured randomly) of input face image, random shuffling of coefficients in the eigen vector (extracted using PCA method) obtained after fusion and addition among coefficients in the shuffled eigen vector. On the theoretical basis, it has been proved that the proposed approach provides better template protection against spoof attacks as compared to the existing method. One of the weaknesses of biometrics is that once a biometric data or template is stolen, it is stolen forever and cannot be reissued, or discarded. Thus template security has become very critical in these systems. The proposed scheme provides new measures (shuffling and addition) for template protection by giving the ability to discard the lost template and reissue a new one.

VII. References

1) Bori Toth, ―Biometric Liveness Detection", Information Security Bulletin, October 2005, Volume 10, pages 291-297.
2) International Biometric Group. Liveness detection in biometric systems, 2003. White paper. Available at http://www.biometricgroup.com/reports/public/reports/liveness.html.
3) A. K. Jain, A. Ross, and S. Prabhakar, ―An introduction to biometric recognition," IEEE Trans. on Circuits and Systems for Video Technology, vol. 14, pp. 4–20, Jan 2004.
4) Arun Ross and Anil K. Jain, ―Multimodal biometrics: An overview", appeared in Proc. of 12th European Signal Processing Conference (EUSIPCO), (Vienna, Austria), pp. 1221-1224, September 2004.
5) Arun Ross, ―An Introduction to Multibiometrics", EUSIPCO, 2007.
6) A. Ross, K. Nandakumar, and A. K. Jain, Handbook of Multibiometrics, New York, Springer, 2006.
7) B. Schneier, ―The uses and abuses of biometrics", Communications of the ACM, vol. 42, no. 8, pp. 136, Aug. 1999.
8) A. K. Jain, K. Nandakumar and A. Nagar, ―Biometric Template Security", EURASIP Journal on Advances in Signal Processing, January 2008.
9) Lu, X., Wang, Y. & Jain, A.K. (2003). Combining Classifiers for Face Recognition, In IEEE Conference on Multimedia & Expo, Vol. 3, pp. 13-16.
10) L. Hong and A.K Jain, ―Integrating faces and fingerprint for personal identification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 12, pp. 1295–1307, 1998.
11) R. Brunelli and D. Falavigna, ―Person identification using multiple cues," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 10, pp. 955–966, Oct. 1995.
12) [12] Xiaoguang Lu, ―Image Analysis for Face Recognition – A brief survey", Dept. of Computer Science & Engineering, Michigan State University, personal notes, May 2003.
13) L. Wiskott, J.M. Fellous, N. Kruger, and C. von der Malsburg, ―Face recognition by elastic bunch graph matching," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775–779, 1997.
14) G.J. Edwards, T.F. Cootes, and C.J. Taylor, ―Face recognition using active appearance models," in Proc. European Conference on Computer Vision, 1998, vol. 2, pp. 581–695.
15) V. Blanz and T. Vetter, ―A morphable model for the synthesis of 3D faces," in Proc. ACM SIGGRAPH, Mar. 1999, pp. 187–194.
16) U. Uludag, S. Pankanti, S. Prabhakar, and A. K. Jain, ―Biometric Cryptosystems: Issues and Challenges," vol. 92, no. 6, pp. 948–960, June 2004.
17) Gang Pan, Zhaohui Wu and Lin sun, ―Liveness Detection for Face Recognition", Recent Advances in Face Recognition, pages 109-123, December 2008, I-Tech, Vienna, Austria.
18) Jiangwei Li, Yunhong Wang, Tieniu Tan, A.K. Jain, ―Live Face Detection Based on the Analysis of Fourier Spectra", Biometric Technology for Human Identification, Proceedings of. SPIE, Vol. 5404.
19) Y. Vijaya Lata, Chandra Kiran Bharadwaj Tungathurthi, H. Ram Mohan Rao, A. Govardhan, L. P. Reddy, ―Facial Recognition using Eigenfaces by PCA", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.
20) L. Sirovich and M. Kirby, ―Low-dimensional procedure for the characterization of human faces", Journal of the Optical Society of America A 4: 519–524, 1987.
21) M.Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, March 1991.

QRS Wave Detection Using Multiresolution Analysis

GJCST Computing Classification
G.1.2, I.5.4, J.3

S.Karpagachelvi[1] Dr.M.Arthanari, Prof. & Head[2] M.Sivakumar[3]

***Abstract*-Te electrocardiogram (ECG or EKG) is basically a diagnostic tool that measures and records the electrical signal by comparing the activity of heart. It is most commonly used to perform cardiac test, since it acts as screening tool for cardiac abnormalities. This is necessary because no single point provides a complete picture of what is going on in the heart. It mainly comprises of PQRS&T wave by showing corresponding time and frequency. PQRST key feature detector is based on wavelet transform which robust to time varying and noise. It will analyze the waveform including noise purification, sample design of digital ECG. R peak is mainly used for detection. In this work, we have developed an electrocardiogram (ECG) feature extraction system based on the multi-resolution wavelet transform. It mainly includes two stages. In the first stage, algorithm is quoted by using discrete wavelet transform for de-noise the signal. In second step multiresolution is done for QRS complex detection. The proposed schemes were mostly based on Fuzzy Logic Methods, Artificial Neural Networks (ANN), Genetic Algorithm (GA), Support Vector Machines (SVM), and other Signal Analysis techniques.**

Keywords-Cardiac Cycle, ECG signal, P-QRS-T waves, Feature Extraction, Haar wavelets.

I. INTRODUCTION

The investigation of the ECG has been extensively used for diagnosing many cardiac diseases. The ECG is a realistic record of the direction and magnitude of the electrical commotion that is generated by depolarization and re-polarization of the atria and ventricles. One cardiac cycle in an ECG signal consists of the P-QRS-T waves. Figure 1 shows a sample ECG signal. The majority of the clinically useful information in the ECG is originated in the intervals and amplitudes defined by its features (characteristic wave peaks and time durations). The improvement of precise and rapid methods for automatic ECG feature extraction is of chief importance, particularly for the examination of long recordings [1].

The ECG feature extraction system provides fundamental features (amplitudes and intervals) to be used in subsequent automatic analysis. In recent times, a number of techniques have been proposed to detect these features [2] [3] [4]. The previously proposed method of ECG signal analysis was based on time domain method. But this is not always adequate to study all the features of ECG signals. Therefore the frequency representation of a signal is required. The deviations in the normal electrical patterns indicate various cardiac disorders. Cardiac cells, in the normal state are electrically polarized [5].

ECG is essentially responsible for patient monitoring and diagnosis. The extracted feature from the ECG signal plays a vital in diagnosing the cardiac disease. The development of accurate and quick methods for automatic ECG feature extraction is of major importance. Therefore it is necessary that the feature extraction system performs accurately. The purpose of feature extraction is to find as few properties as possible within ECG signal that would allow successful abnormality detection and efficient prognosis.

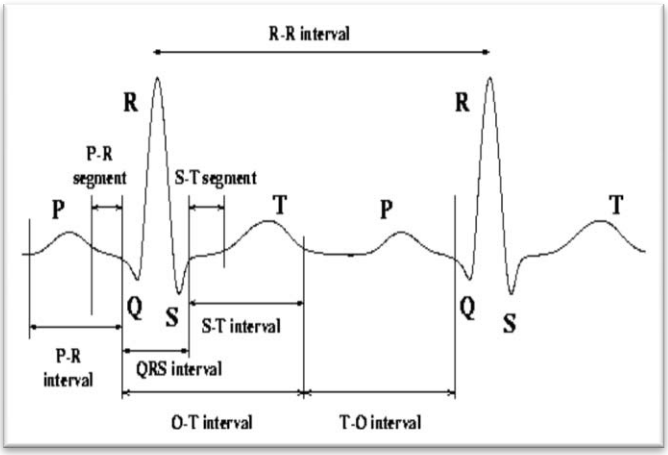


Figure.1 A Sample ECG Signal showing P-QRS-T Wave

recent year, several research and algorithm have been developed for the exertion of analyzing and classifying the ECG signal. The classifying method which have been proposed during the last decade and under evaluation includes digital signal analysis, Fuzzy Logic methods, Artificial Neural Network, Hidden Markov Model, Genetic Algorithm, Support Vector Machines, Self-Organizing Map, Bayesian and other method with each approach exhibiting its own advantages and disadvantages. In this work, we have developed an electrocardiogram (ECG) feature extraction system based on the multi-resolution wavelet transform using haar coefficients and also provide an over view on various techniques and transformations used for extracting the feature from ECG signal. This paper is structured as follows. Section 2 discusses the related work that was earlier proposed in literature for ECG feature extraction. Section 3

About-[1]Doctoral Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India.(email;karpagachelvis@yahoo.com)
About-[2]Dept. of Computer Science and Engineering,Tejaa Shakthi Institute of Technology for Women Coimbatore- 641 659, Tamilnadu, India (email: arthanarimsvc@gmail.com)
About-3 Doctoral Research Scholar Anna University – Coimbatore Tamilnadu, India(email;sivala@gmail.com)

gives a description of the DWT based ECG feature detection algorithm and Section 4 concludes the paper with fewer discussions

II. Related work

ECG feature extraction has been studied from early time and lots of advanced techniques as well as transformations have been proposed for accurate and fast ECG feature extraction. This section of the paper discusses various techniques and transformations proposed earlier in literature for extracting feature from ECG.

A novel approach for ECG feature extraction was put forth by Castro et al. in [6]. Their proposed paper present an algorithm, based on the wavelet transform, for feature extraction from an electrocardiograph (ECG) signal and recognition of abnormal heartbeats. Since wavelet transforms can be localized both in the frequency and time domains. They developed a method for choosing an optimal mother wavelet from a set of orthogonal and bi-orthogonal wavelet filter bank by means of the best correlation with the ECG signal. The coefficients, approximations of the last scale level and the details of the all levels, are used for the ECG analyzed. They divided the coefficients of each cycle into three segments that are related to P-wave, QRS complex, and T-wave. The summation of the values from these segments provided the feature vectors of single cycles.

Mahmoodabadi et al. in [1] described an approach for ECG feature extraction which utilizes Daubechies Wavelets transform. They had developed and evaluated an electrocardiogram (ECG) feature extraction system based on the multi-resolution wavelet transform. The ECG signals from Modified Lead II (MLII) were chosen for processing. The wavelet filter with scaling function further intimately similar to the shape of the ECG signal achieved better detection. The foremost step of their approach was to de-noise the ECG signal by removing the equivalent wavelet coefficients at higher scales. Then, QRS complexes are detected and each one complex is used to trace the peaks of the individual waves, including onsets and offsets of the P and T waves which are present in one cardiac cycle.

A feature extraction method using Discrete Wavelet Transform (DWT) was proposed by Emran et al. in [7]. They used a discrete wavelet transform (DWT) to extract the relevant information from the ECG input data in order to perform the classification task. Their proposed work includes the following modules data acquisition, pre-processing beat detection, feature extraction and classification. In the feature extraction module the Wavelet Transform (DWT) is designed to address the problem of non-stationary ECG signals. It was derived from a single generating function called the mother wavelet by translation and dilation operations. Using DWT in feature extraction may lead to an optimal frequency resolution in all frequency ranges as it has a varying window size, broad at lower frequencies, and narrow at higher frequencies. The DWT characterization will deliver the stable features to the morphology variations of the ECG waveforms.

Tayel and Bouridy together in [8] put forth a technique for ECG image classification by extracting their feature using wavelet transformation and neural networks. Features are extracted from wavelet decomposition of the ECG images intensity. The obtained ECG features are then further processed using artificial neural networks. The features are: mean, median, maximum, minimum, range, standard deviation, variance, and mean absolute deviation. The introduced ANN was trained by the main features of the 63 ECG images of different diseases.

An algorithm was presented by Chouhan and Mehta in [9] for detection of QRS complexities. The recognition of QRS-complexes forms the origin for more or less all automated ECG analysis algorithms. The presented algorithm utilizes a modified definition of slope, of ECG signal, as the feature for detection of QRS. A succession of transformations of the filtered and baseline drift corrected ECG signal is used for mining of a new modified slope-feature. In the presented algorithm, filtering procedure based on moving averages [15] provides smooth spike-free ECG signal, which is appropriate for slope feature extraction. The foremost step is to extort slope feature from the filtered and drift corrected ECG signal, by processing and transforming it, in such a way that the extracted feature signal is significantly enhanced in QRS region and suppressed in non-QRS region.

Xu et al. in [10] described an algorithm using Slope Vector Waveform (SVW) for ECG QRS complex detection and RR interval evaluation. In their proposed method variable stage differentiation is used to achieve the desired slope vectors for feature extraction, and the non-linear amplification is used to get better of the signal-to-noise ratio. The method allows for a fast and accurate search of the R location, QRS complex duration, and RR interval and yields excellent ECG feature extraction results. In order to get QRS durations, the feature extraction rules are needed.

A modified combined wavelet transforms technique was developed by Saxena et al. in [11]. The technique has been developed to analyze multi lead electrocardiogram signals for cardiac disease diagnostics. Two wavelets have been used, i.e. a quadratic spline wavelet (QSWT) for QRS detection and the Daubechies six coefficient (DU6) wavelet for P and T detection. A procedure has been evolved using electrocardiogram parameters with a point scoring system for diagnosis of various cardiac diseases. The consistency and reliability of the identified and measured parameters were confirmed when both the diagnostic criteria gave the same results. Table 1 shows the comparison of different ECG signal feature extraction techniques.

Fatemian et al.[12] proposed an approach for ECG feature extraction. They suggested a new wavelet based framework for automatic analysis of single lead electrocardiogram (ECG) for application in human recognition. Their system utilized a robust preprocessing stage, which enables it to handle noise and outliers. This facilitates it to be directly applied on the raw ECG signal. In addition the proposed system is capable of managing ECGs regardless of the heart rate (HR) which renders making presumptions on the individual's stress level unnecessary. The substantial reduction of the template gallery size decreases the storage requirements of the system appreciably. Additionally, the categorization process is speeded up by eliminating the need for dimensionality reduction techniques such as PCA or LDA. Their experimental results revealed the fact that the proposed technique out performed other conventional methods of ECG feature extraction.

III. Description of algorithm

A. *Wavelet Selection*

The large number of known wavelet families and functions provides a rich space in which to search for a wavelet which will very efficiently represent a signal of interest in a large variety of applications. Wavelet families include Biorthogonal, Coiflet, Harr, Symmlet, Daubechies wavelets, etc. There is no absolute way to choose a certain wavelet. The choice of the wavelet function depends on the application. The Haar wavelet algorithm has the advantage of being simple to compute and easy to understand. In the present work Haar wavelet is chosen. Savitzky Golay filtering is used to smooth the signal. To identify the onsets and offsets of the wave , the wave is made to zero base. To obtain the wavelet analysis, we used the Matlab program, which contains a very good —wavelet toolbox”. First the considered signal was decomposed using Haar wavelet of the order of 1-5 has been evaluated. One of the key criteria of a good mother wavelet is its ability to fully reconstruct the signal from the wavelet decompositions. The fig 2 shows the decomposed signal. The high frequency components of the ECG signal decreases as lower details are removed from the original signal. As the lower details are removed, the signal becomes smoother and the noises disappears since noises are marked by high frequency components picked up along the ways of transmission. This is the contribution of the discrete wavelet transform where noise filtration is performed implicitly.

B. *Peaks identification*

In order to detect the peaks, specific details of the signal were selected. R peaks are the Largest amplitude points which are greater than threshold points are located in the wave. Those maxima points are stored and the R-R interval is determined. Their mean value is found which is used to find the portion of the single wave. A Q and S peak occurs about the R peak with in 0.1second. Calculating the distance from zero point or close zero left side of R peak within the threshold limit denotes Q peak. The onset is the beginning of the Q wave (or R-wave if the Q-wave is missing) and the offset is the ending of the S-wave (or R-wave if the S wave is missing). Normally, the onset of the QRS complex contains the high-frequency components, which are detected at finer scales. Calculating the distance from zero point or close zero right side of R peak within the threshold limit denotes Q peak.

C. *Results*

The algorithm presented in this section is applied directly at one run over the whole digitized ECG signal which are saved as data files provided by Physionet.

QRS recognition is shown in Figure 3.

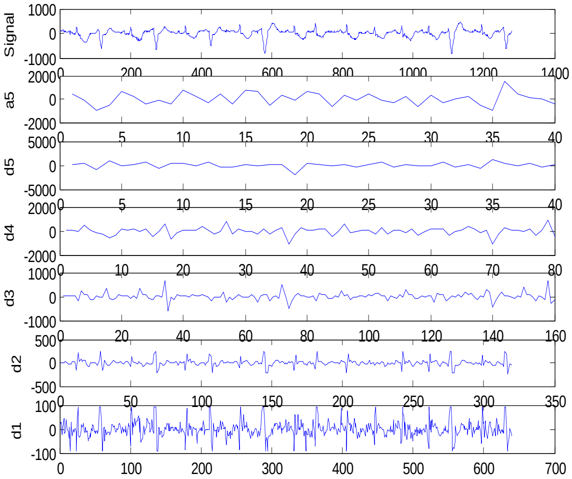


Fig.2. Multiresolution decomposition of ECG signal from 801.dat file

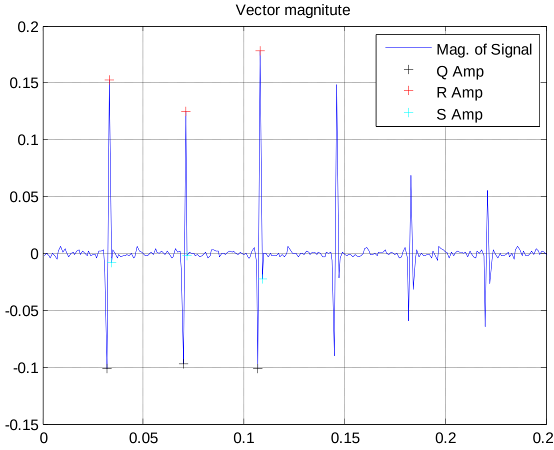


Fig.3. Multiresolution process of wavelet-based peak Detection in 801.dat file

IV. Conclusion

In this paper, QRS key feature elements detection algorithm based on multi resolution analysis was proposed. The performance of the peak detection was examined by testing the algorithm on data standardized MIT-BIH database. The DWT based QRS detector performs well with standard techniques. Thus, the primary advantages of the DWT over existing techniques are noise removal and ability to process the time varying ECG data. In this work we pointed out the advantage of using wavelet transform associated with a threshold

strategy. Further, the possibility of detecting positions of QRS complexes in ECG signals is investigated and a simple detection algorithm is proposed. The main advantage of this kind of detection is less time consuming analysis for long time ECG signal. The QRS detection in the ECG signal is explained with screen shots. The future work mainly concentrates on improving the proposed algorithm for various QRS waves of different patients. Moreover additional statistical data will be utilized for evaluating the performance of an algorithm in ECG signal feature detection. Improving the accuracy of diagnosing the cardiac disease at the earliest is necessary in the case of patient monitoring system. Therefore our future work also has an eye on improvement in diagnosing the cardiac disease.

V. References

1) S. Z. Mahmoodabadi, A. Ahmadian, and M. D. Abolhasani, ―ECG Feature Extraction using Daubechies Wavelets," Proceedings of the fifth IASTED International conference on Visualization, Imaging and Image Processing, pp. 343-348, 2005.
2) Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna, ―A Wavelet-Based ECG Delineator: Evaluation on Standard Databases," IEEE Transactions on Biomedical Engineering Vol. 51, No. 4, pp. 570-581, 2004.
3) Krishna Prasad and J. S. Sahambi, ―Classification of ECG Arrhythmias using Multi-Resolution Analysis and Neural Networks," IEEE Transactions on Biomedical Engineering, vol. 1, pp. 227-231, 2003.
4) Cuiwei Li, Chongxun Zheng, and Changfeng Tai, ―Detection of ECG Characteristic Points using Wavelet Transforms," IEEE Transactions on Biomedical Engineering, Vol. 42, No. 1, pp. 21-28, 1995.
5) Saritha, V. Sukanya, and Y. Narasimha Murthy, ―ECG Signal Analysis Using Wavelet Transforms," Bulgarian Journal of Physics, vol. 35, pp. 68-77, 2008.
6) B. Castro, D. Kogan, and A. B. Geva, ―ECG feature extraction using optimal mother wavelet," The 21st IEEE Convention of the Electrical and Electronic Engineers in Israel, pp. 346-350, 2000.
7) Emran M. Tamil, Nor Hafeezah Kamarudin, Rosli Salleh, M. Yamani Idna Idris, Noorzaily M.Noor, and Azmi Mohd Tamil, ―Heartbeat Electrocardiogram (ECG) Signal Feature Extraction Using Discrete Wavelet Transforms (DWT)."
8) Mazhar B. Tayel, and Mohamed E. El-Bouridy, ―ECG Images Classification Using Feature Extraction Based On Wavelet Transformation And Neural Network," ICGST, International Conference on AIML, June 2006.
9) V. S. Chouhan, and S. S. Mehta, ―Detection of QRS Complexes in 12-lead ECG using Adaptive Quantized Threshold," IJCSNS International Journal of Computer Science and Network Security, vol. 8, no. 1, 2008.
10) Xiaomin Xu, and Ying Liu, ―ECG QRS Complex Detection Using Slope Vector Waveform (SVW) Algorithm," Proceedings of the 26th Annual International Conference of the IEEE EMBS, pp. 3597-3600, 2004.
11) S. C. Saxena, V. Kumar, and S. T. Hamde, ―Feature extraction from ECG signals using wavelet transforms for disease diagnostics," International Journal of Systems Science, vol. 33, no. 13, pp. 1073-1085, 2002.
12) S. Z. Fatemian, and D. Hatzinakos, ―Anew ECG feature extractor for biometric recognition," 16th International Conference on Digital Signal Processing, pp. 1-6, 2009.

A Review on Data Clustering Algorithms for Mixed Data

GJCST Computing Classification
H.3.3, H.2.8

D. Hari Prasad[1] Dr. M. Punithavalli[2]

***Abstract*-Clustering is the unsupervised classification of patterns into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis. In general, clustering is a method of dividing the data into groups of similar objects. One of significant research areas in data mining is to develop methods to modernize knowledge by using the existing knowledge, since it can generally augment mining efficiency, especially for very bulky database. Data mining uncovers hidden, previously unknown, and potentially useful information from large amounts of data. This paper presents a general survey of various clustering algorithms. In addition, the paper also describes the efficiency of Self-Organized Map (SOM) algorithm in enhancing the mixed data clustering**

Keywords-Data Clustering, Data Mining, Mixed Data Clustering, Self-Organized Map algorithm.

I. Introduction

Clustering is one of the standard workhorse techniques in the field of data mining. Its intention is to systematize a dataset into a set of groups, or clusters, which contain –similar” data items, as measured by some distance function. The major applications of clustering include document categorization, scientific data analysis, and customer/market segmentation. Data clustering has been considered as a primary data mining method for knowledge discovery. Clustering using Gaussian mixture models is also extensively employed for exploratory data analysis. The six sequential, iterative steps of Data mining processes are: 1) problem definition; 2) data acquisition; 3) data preprocessing and survey; 4) data modeling; 5) evaluation; 6) knowledge deployment [1]. The purpose of survey before data preprocessing is to gain insight knowledge into the data possibilities and problems to determine whether the data are sufficient. Moreover the survey assists us to select the proper preprocessing and modeling tools. Typically, several different data sets and preprocessing strategies need to be considered. For this reason, efficient visualizations and summarizations are essential.

Primarily the focus must be on clustering since they are important characterizations of data. The clustering method implemented should be fast, robust, and visually efficient. In the case of clustering Q means, the foremost step is partitioning a data set into a set of clusters Q_i, where i = 1 C. Data clustering techniques are gaining escalating reputation over traditional central grouping techniques, which are centered on the conception of –feature” (see e.g. [2], [3]). Several data clustering techniques have been put forth by researchers to assist in the development of knowledge.

Fuzzy clustering [4] is a simplification of crisp clustering where each sample has a varying degree of membership in all clusters. In many real-world applications, in fact, a feasible feature-based description of objects might be difficult to obtain or inefficient for learning purposes while, on the other hand, it is often possible to obtain a measure of the similarity or dissimilarity between objects. Among the central algorithmic procedures for perceptual organization are clustering principles like generalized k-means methods or clustering methods for proximity data [15].

The remainder of this paper is organized as follows section II describes the background study that is related to clustering algorithms proposed earlier, section III explains the challenging problems and areas of research and section IV concludes the paper with fewer discussions.

II. Background Study

A wealth of clustering techniques had been described in the literature. This section of the paper presents an overview on these clustering algorithms put forth by various researchers. In general, major clustering methods can be classified into five categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods.

A. Clustering of the Self-Organizing Map

A novel method [1] was put forth by Juha Vesanto and Esa Alhoniemi for clustering of Self-Organizing Map. According to the method proposed in this paper the clustering is carried out using a two-level approach, where the data set is first clustered using the SOM, and then, the SOM is clustered. The purpose of this paper was to evaluate if the data abstraction created by the SOM could be employed in clustering of data. The most imperative advantage of this procedure is that computational load decreases noticeably, making it possible to cluster large data sets and to consider several different preprocessing strategies in a restricted time. Obviously, the approach is applicable only if the clusters found using the SOM are analogous to those of the original data.

B. Kernel-Based Clustering

Mark Girolami presents a Mercer Kernel-Based Clustering [5] algorithm in Feature Space. This paper presents a method for both the unsupervised partitioning of a sample of data and the estimation of the possible number of inherent

About-[1]Senior Lecturer, Department of Computer Applications, Sri Ramakrishna Institute of Technology, Coimbatore, India.

About-[2]Director, Department of Computer Science, Sri Ramakrishna Arts College for Women, Coimbatore, India.

clusters which generate the data. This work utilizes the perception that performing a nonlinear data transformation into some high dimensional feature space increases the probability of the linear separability of the patterns within the transformed space and therefore simplifies the associated data structure. In this case, the eigenvectors of a kernel matrix which defines the implicit mapping provides a means to estimate the number of clusters inherent within the data and a computationally simple iterative procedure is presented for the subsequent feature space partitioning of the data.

C. Grouping of Smooth Curves and Texture Segmentation using path-based clustering

A Path-Based Clustering algorithm [6] was described by Fischer and Buhmann for grouping of smooth curves and texture segmentation. This paper proposed a new grouping approach referred to as Path-Based Clustering [7], which measures local homogeneity rather than global similarity of objects. The new Path-Based Clustering method defines a connectedness criterion, which groups objects together if they are connected by a sequence of intermediate objects. Moreover an efficient agglomerative algorithm is proposed to minimize the Path-Based Clustering cost function. This approach utilizes a bootstrap resampling scheme to measure the reliability of the grouping results.

D. Bagging for Path-Based Clustering

Fischer and Buhmann present bagging for path-based clustering [8]. A resampling scheme for clustering with similarity to bootstrap aggregation (bagging) is presented in this paper. This aggregation (Bagging) is used to develop the quality of path-based clustering, a data clustering method that can extort stretched out structures from data in a noise stout way. In order to increase the reliability of clustering solutions, a stochastic resampling method is developed to deduce accord clusters. Moreover this paper also evaluates the quality of path-based clustering with resampling on a large image dataset of human segmentations.

E. Isoperimetric Graph Partitioning for Data Clustering

Leo Grady and Eric L. Schwartz together proposed an approach known as Isoperimetric Graph Partitioning for Data Clustering and Image Segmentation [9]. This paper, adopts a different approach, based on finding partitions with a small isoperimetric constant in an image graph. The algorithm described in this paper generates high quality segmentations and data clusters of spectral methods, but with improved speed and stability. The term ―partition" in this paper refers to the assignment of each node in the vertex set into two (not necessarily equal) parts. Graph partitioning has been strongly influenced by properties of a combinatorial formulation of the classic isoperimetric problem: For a fixed area, find the region with minimum perimeter.

F. Improving Classification Decisions by Multiple Knowledge

The new approach to combine multiple sets of rules for text categorization using Dempster‗s rule of combination [10] was described by Yaxin Bi et al. A boosting-like technique for generating multiple sets of rules based on rough set theory and model classification decisions from multiple sets of rules as pieces of evidence which can be combined by Dempster‗s rule of combination is developed in this approach. This approach is employed to set of benchmark data collection, both individually and in combination. The experimental results show that the performance of the best combination of the multiple sets of rules on the benchmark data is significantly better than that of the best single set of rules.

G. Clustering Algorithm for Data Mining

Zhijie Xu et al. expressed a Modified Clustering Algorithm for Data Mining [11]. This paper describes a clustering method for unsupervised classification of objects in large data sets. The new methodology particularly combines the simulating annealing algorithm with CLARANS (clustering Large Application based upon Randomized Search) in order to cluster large data sets efficiently. The parameter T is used to control the process of clustering. In every step of the search, if the cost of the neighbor is less than the current, set the current to the neighbor. Otherwise, accept the neighbor with the probability of exp (-(Scost-currentcost)/T).

H. Dominant Sets and Pairwise Clustering

A graph-theoretic approach [12] for Pairwise data clustering was developed by Massimiliano Pavan and Marcello Pelillo. A correspondence is established between dominant sets and the extrema of a quadratic form over the standard simplex, thereby allowing the use of straightforward and easily implementable continuous optimization techniques from evolutionary game theory. In order to study the robustness of the approach against random noise in the background, the level of clutter is allowed to vary, starting from 100 to 1,000 points. Extensions of the approach presented in this paper involving hierarchical data partitioning and out of-sample extensions of dominant-set clusters can be found in [13], and [14], respectively.

I. A Conceptual Clustering Algorithm

Biswas et al. in [17] put forth a conceptual clustering algorithm for data mining. Their paper described an unsupervised discovery method with biases geared toward partitioning objects into clusters that improve interpretability. Their algorithm, ITERATE, employs: (i) a data ordering scheme and (ii) an iterative redistribution operator to produce maximally cohesive and distinct clusters. The important task here is interpretation of the generated patterns, and this is best addressed by creating groups of data that demonstrate cohesiveness within but clear distinctions between the groups. In clustering schemes, data objects are represented as vectors of feature-value pairs.

Features represent properties of an object that are relevant to the problem-solving task. Distinctness or inter-class dissimilarity was measured by an average of the variance of the distribution match between clusters. Additionally, their empirical results demonstrated the properties of the discovery algorithm, and its applications to problem solving.

J. The New K-Windows Algorithm for Improving the K-Means Clustering Algorithm

The new K-windows algorithm for improving the K-means clustering algorithm was described by Vrahatis et al. in [18]. The process of partitioning a large set of patterns into disjoint and homogeneous clusters is fundamental in knowledge acquisition. It is called Clustering in the literature and it is applied in various fields including data mining, statistical data analysis, compression and vector quantization. The k-means is a very popular algorithm and one of the best for implementing the clustering process. The k-means has a time complexity that is dominated by the product of the number of patterns, the number of clusters, and the number of iterations. Also, it often converges to a local minimum. In their paper, they presented an improvement of the k-means clustering algorithm, aiming at a better time complexity and partitioning accuracy. Moreover, their approach reduces the number of patterns that are needed to be examined for similarity using a windowing technique. The latter is based on well known spatial data structures, namely the range tree, which allows fast range searches.

K. A Spectral-based Clustering Algorithm

Abdu et al. in [19] presented a novel spectral-based algorithm for clustering categorical data that combines attribute relationship and dimension reduction techniques found in Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI). The new algorithm uses data summaries that consist of attribute occurrence and co-occurrence frequencies to create a set of vectors each of which represents a cluster. They referred to these vectors as ―candidate cluster representatives.” The algorithm also uses spectral decomposition of the data summaries matrix to project and cluster the data objects in a reduced space. They referred to the algorithm as SCCADDS (Spectral-based Clustering algorithm for CAtegorical Data using Data Summaries). SCCADDS differs from other spectral clustering algorithms in several key respects. Initially, the algorithm uses the feature categories similarity matrix instead of the data object similarity matrix (as is the case with most spectral algorithms that find the normalized cut of a graph of nodes of data objects). SCCADDS scales well for large datasets. Second, non-recursive spectral-based clustering algorithms characteristically necessitate K-means or some other iterative clustering method after the data objects have been projected into a reduced space. SCCADDS clusters the data objects directly by comparing them to candidate cluster representatives without the need for an iterative clustering method. Third, unlike standard spectral-based algorithms, the complexity of SCCADDS is linear in terms of the number of data objects. Results on datasets widely used to test categorical clustering algorithms show that SCCADDS produces clusters that are consistent with those produced by existing algorithms, while avoiding the computation of the spectra of large matrices and problems inherent in methods that employ the K-means type algorithms

L. A New Supervised Clustering Algorithm

A new supervised clustering algorithm was projected by Li et al. in [20]. They suggested their algorithm for data set with mixed attributes. Because of the complexity of data set with mixed attributes, the conventional clustering algorithms appropriate for this kind of dataset are not many and the result of clustering is not good. K-prototype clustering is one of the most commonly used methods in data mining for this kind of data. They borrowed the ideas from the multiple classifiers combing technology, use k- prototype as the basis clustering algorithm in order to design a multi-level clustering ensemble algorithm in the paper, which adoptively selects attributes for re-clustering. Comparison experiments on Adult data set from UCI machine learning data repository show very competitive results and the proposed method is suitable for data editing.

M. An Efficient Clustering Algorithm for mixed type attributes in Large Dataset

Jian et al. in [21] proposed an efficient algorithm for clustering mixed type attributes in large dataset. Clustering is a extensively used technique in data mining. At present there exist many clustering algorithms, but most existing clustering algorithms either are restricted to handle the single attribute or can handle both data types but are not competent when clustering large data sets. Few algorithms can do both well. In this article, they proposed a clustering algorithm that can handle large datasets with mixed type of attributes. They first used CF*tree (just like CF-tree in BIRCH) to pre-cluster datasets. After that the dense regions are stored in leaf nodes, and then they looked every dense region as a single point and used the ameliorated k-prototype to cluster such dense regions. Experimental results showed that this algorithm is very efficient in clustering large datasets with mixed type of attributes.

N. A Robust and Scalable Clustering Algorithm

A robust and scalable clustering algorithm was put forth by Chiu et al. in [22]. They employed this clustering algorithm for mixed type attributes in large database environment. In their paper, they proposed a distance measure that enables clustering data with both continuous and categorical attributes. This distance measure is derived from a probabilistic model that the distance between two clusters is equivalent to the decrease in log-likelihood function as a result of merging. Calculation of this measure is memory efficient as it depends only on the merging cluster pair and not on all the other clusters. The algorithm is implemented in the commercial data mining tool Clementine 6.0 which supports the PMML standard of data mining model deployment. For data with mixed type of attributes, their experimental results confirmed that the algorithm not only generates better quality clusters than the traditional k-means algorithms, but also exhibits good scalability properties and is able to identify the underlying number of clusters in the data correctly

O. Clustering Algorithm for Network Intrusion Detection system

Panda et al. in [23] described some clustering algorithms such as K-Means and Fuzzy c-Means for network intrusion detection. The objective of intrusion detection is to construct a system which would automatically scan network activity and detect such intrusion attacks. They built a system which created clusters from its input data, then automatically labeled clusters as containing either normal or anomalous data instances, and finally used these clusters to classify network data instances as either normal or anomalous. In their paper, they intended to propose a fuzzy c-means clustering technique which is capable of clustering the most suitable number of clusters based on objective function. Both the training and testing was done using 10% KDDCup‘99 data, which is a very well-liked and broadly used intrusion attack dataset.

P. Clustering Algorithm-based on Quantum Games

A new clustering algorithm based on quantum games was projected by Li et al. in [24]. Mammoth successes have been made by quantum algorithms during the last decade. In their paper, they combined the quantum game with the problem of data clustering, and then they developed a quantum-game-based clustering algorithm, in which data points in a dataset are considered as players who can make decisions and implement quantum strategies in quantum games. After each round of a quantum game, each player's expected payoff is calculated. Soon after, he uses a link-removing-and-rewiring (LRR) function to change his neighbors and regulate the strength of links connecting to them in order to maximize his payoff. Further, algorithms are discussed and analyzed in two cases of strategies, two payoff matrixes and two LRR functions. Accordingly, the simulation results have demonstrated that data points in datasets are clustered reasonably and efficiently, and the clustering algorithms have fast rates of convergence. Furthermore, the comparison with other algorithms also provides an indication of the effectiveness of the proposed approach

Q. A GA-based Clustering Algorithm

Jie Li et al. in [25] proposed a GA-based clustering algorithm for large data sets with mixed and numeric and categorical values. In the field of data mining, it is frequently encountered to execute cluster analysis on large data sets with mixed numeric and categorical values. However, most existing clustering algorithms are only competent for the numeric data rather than the mixed data set. For this reason, their paper presented a novel clustering algorithm for these mixed data sets by modifying the common cost function, trace of the within cluster dispersion matrix. The genetic algorithm (GA) is used to optimize the new cost function to obtain valid clustering result. Experimental result illustrates that the GA-based new clustering algorithm is reasonable for the large data sets with mixed numeric and categorical values.

III. Challenging Problems And Areas Of Research

The algorithms proposed by researchers discussed in section II of this paper have their own advantages and limitations. The main requirements that a clustering algorithm should satisfy are: scalability, dealing with different types of attributes, discovering clusters with arbitrary shape, minimal requirements for domain knowledge to determine input parameters, ability to deal with noise and outliers, insensitivity to order of input records, high dimensionality, interpretability and usability. A number of problems are associated with conventional clustering algorithms. A few among them are current clustering techniques do not address all the requirements adequately (and concurrently), dealing with large number of dimensions and large number of data items can be problematic because of time complexity, the effectiveness of the method depends on the definition of ―distance” (for distance-based clustering), if an obvious distance measure doesn‘t exist, then one must ―define” it, which is not always easy, especially in multi-dimensional spaces, the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways [16]. A lot of algorithms for clustering data have been developed in recent decades, nonetheless, they all visage a major challenge in scaling up to very large database sizes, an accelerating development brought on by advances in computer technology, the Internet, and electronic commerce. The mainly focused research area is Clustering of mixed data. A clustering Q means partitioning a data set into a set of clusters Qi, where i = 1… C. In crisp clustering, each data sample belongs to exactly one cluster. Clustering algorithms may be classified as Exclusive Clustering, Overlapping Clustering, Hierarchical Clustering, and Probabilistic Clustering. Clustering objects into separated groups is an important topic in exploratory data analysis and pattern recognition. Many clustering techniques group the data objects together to ―compact” clusters with the explicit or implicit assumption that all objects within one group are either mutually similar to each other or they are similar with respect to a common representative or Centroid. Clustering can also be based on mixture models [1]. In this approach, the data are assumed to be generated by several parameterized distributions (typically Gaussians). Distribution parameters are estimated using, for example, the expectation-maximation algorithm. Data points are assigned to different clusters based on their probabilities in the distributions. The implementation of clustering algorithms to mixed data is one of the challenging issues

IV. Conclusion

This proposed paper describes various algorithms presented by researchers for data clustering. Most of the real time applications need clustering of data. This data clustering can be implemented to mixed data which is the combination of numeric and strings. The clustering algorithm proposed in literature may have its own advantages and limitations. Developing an algorithm that meets all the requirements of the system is tangible. Different clustering algorithms like k-means, path-based clustering, clustering of self organized map are used widely for real world applications. The future work mainly concentrates on developing a clustering algorithm that meets all the requirements. Moreover, the future enhancement vision to develop a clustering algorithm that performs significantly well for mixed data set

V. REFERENCES

1) Juha Vesanto and Esa Alhoniemi, —Clustering of Self-Organizing Map," IEEE Transactions on Neural Networks, vol. 11, no. 3, May 2000, pp. 586-600.
2) J. Shi and J. Malik, —Normalized Cuts and Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, Aug. 2000.
3) Y. Gdalyahu, D. Weinshall, and M. Werman, —Self-Organization in Vision: Stochastic Clustering for Image Segmentation, Perceptual Grouping, and Image Database Organization," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1053-1074, Oct. 2001.
4) J. C. Bezdek and S. K. Pal, Eds., —Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data," New York: IEEE, 1992.
5) Mark Girolami, —Mercer Kernel-based Clustering in Feature space," IEEE Transactions on Neural Networks, vol. 13, no. 3, May 2002.
6) Bernd Fischer, and J. M. Buhmann, —Path-Based Clustering for Grouping of Smooth Curves and Texture Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 4, April 2003.
7) Fischer, T. Zoller, and J.M. Buhmann, —Path Based Pair wise Data Clustering with Application to Texture Segmentation," Energy Minimization Methods in Computer Vision and Pattern Recognition, pp. 235-250, LNCS 2134, 2001.
8) Bernd Fischer, and J. M. Buhmann, —Bagging for Path-Based Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, November 2003.
9) Leo Grady and Eric L. Schwartz, —Isoperimetric Graph Partitioning for Data Clustering and Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004.
10) Yaxin Bi, Sally McClean and Terry Anderson, —Improving Classification Decisions by Multiple Knowledge," Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, 2005.
11) Zhijie Xu, Laisheng Wang, Jiancheng Luo and Jianqin Zhang, —AModified Clustering Algorithm Data Mining," IEEE 2005.
12) Massimiliano Pavan and Marcello Pelillo, —Dominant Sets and Pairwise Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, January 2007.
13) M. Pavan and M. Pelillo, —Dominant Sets and Hierarchical Clustering," Proceedings of IEEE International Conference Computer Vision, vol. 1, pp. 362-369, 2003.
14) M. Pavan and M. Pelillo, —Efficient Out-of-Sample Extension of Dominant-Set Clusters," Advances in Neural Information Processing Systems 17,L.K. Saul, Y. Weiss, and L. Bottou, eds., pp. 1057-1064, 2005.
15) J. M. Buhmann, —Data Clustering and Learning," Handbook of Brain Theory and Neural Networks, M. Arbib, ed., pp. 308-312, Bradfort Books/MIT Press, second ed., 2002.
16) A Tutorial on Clustering Algorithms,http://home .dei.polimi.it/matteucc/Clustering/tutorial_html.
17) Gautam Biswas, Jerry B. Weinberg, and Douglas H. Fisher, —ITERATE: A Conceptual Clustering Algorithm for Data Mining," IEEE Transactions on Systems, Man, and Cybernetics, vol. 28, part c, no. 2, pp. 100-111, 1998.
18) M. N. Vrahatis, B. Boutsinas, P. Alevizos, and G. Pavlides, —The New k-Windows Algorithm for Improving the k -Means Clustering Algorithm," Journal of Complexity, Elsevier, vol. 18, no. 1, pp. 375-391, 2002.
19) Eman Abdu, and Douglas Salane, —Aspectral-based clustering algorithm for categorical data using data summaries," International Conference on Knowledge Discovery and Data Mining, ACM, Article no. 2, 2009.
20) Shijin Li, Jing Liu, Yuelong Zhu, and Xiaohua Zhang, —ANew Supervised Clustering Algorithm for Data Set with Mixed Attributes," Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, vol. 2, pp. 844-849, 2007.
21) Jian Yin, Zhi-Fang Tan, Jiang-Tao Ren, and Yi-Qun Chen, —Anefficient clustering algorithm for mixed type attributes in large dataset," Proceedings of 2005 International Conference on Machine Learning and Cybernetics, vol. 3, pp. 1611-1614, 2005.
22) Tom Chiu, DongPing Fang, John Chen, Yao Wang, and Christopher Jeris, —A robust and scalable clustering algorithm for mixed type attributes in large database environment," International Conference on Knowledge Discovery and Data Mining, pp. 263-268, 2001.
23) Mrutyunjaya Panda, and Manas Ranjan Patra, —Some Clustering Algorithms to Enhance the Performance of the Network Intrusion Detection System," Journal of Theoretical and Applied Information Technology, pp. 710-716, 2008.
24) Qiang Li, Yan He, and Jing-ping Jiang, —Anovel clustering algorithm based on quantum games,"

Journal of Physics A: Mathematical and Theoritical, no. 44, 2009.

25) Jie Li, Xinbo Gao, and Li-cheng Jiao, ―AGA-Based Clustering Algorithm for Large Data Sets with Mixed Numeric and Categorical Values," Proceedings of the 5th International Conference on Computational Intelligence and Multimedia Applications, IEEE Computer Society, p. 102, 2003

Optimization Of Shop Floor Operations: Application Of Mrp And Lean Manufacturing Principles

GJCST Computing Classification
J.6, FOR A 091005, 091299

Remy Uche[1]J.A. Onuoha[2]

***Abstract*-This research work is concerned with the optimization of shop floor operations by the application of Material Requirements Planning (MRP) and lean manufacturing principles. The present research covers the involvement of MRP and lean manufacturing techniques in manufacturing environment. The work is intended to decrease cycle time, reduce waste in material movement and inventory, improve the flow of material through improved system layouts and subsequently increase productivity in shop floor environment.**

Keywords-Material Planning, Lean Manufacturing, Scheduling, Production, cycle time.

I. Introduction

Increasing shop floor efficiency through the integration of Material Requirements Planning (MRP) and Lean manufacturing principles has become one of the major concerns of manufacturing companies. In today's complex manufacturing sector, we are confronted to do more with less, and also challenged with new philosophies and concepts that often push or pull us in different directions. A case in point is the ongoing integration of MRP and lean manufacturing principles. MRP systems are frequently condemned as one of the main reasons so many manufacturing companies, are locked into push systems, while lean concepts imply that pull systems are the ideal. Nevertheless, one shouldn't throw one out for the other, as the two can coexist harmoniously and beneficially with a better definition of roles (Steinbrunner, 2004).

According to the American production and control society, MRP constitutes of a set of techniques that use master production schedule, bill of material and inventory data to calculate material requirements. In simple words, MRP is a technique use in determining when to order dependent demand items and how to reschedule orders to adjust for the changing needs. A key question to MRP process is the number of times a company procures inventory within a year. One can readily realize that a high inventory ratio is likely to be conducive to lowering production cost since less capital is tied up to unused inventory.

MRP systems relies on four pieces of information in determining what material should be ordered and when. Namely:

The master production schedule: This describes when each product is scheduled to be manufactured;

Bill of materials: Gives information about the product structure, i.e., parts and raw material units necessary to manufacture one unit of the product of interest;

Production cycle times and material needs at each stage of the production cycle time and Supplier lead times.

The master production schedule and bill of materials indicate what materials should be ordered; the master schedule, production cycle times and supplier lead times jointly determine when orders should be placed.

The Master Production Schedule includes quantities of products to be produced at a given time period.

The Lean Manufacturing is a production method that calls for building products with as few steps and as little work-in-process inventory as possible. It relies on work centres or manufacturing cells that are capable of building multiple products, giving the company the flexibility to produce the exact mix and quantity of products required.

Its fundamental objective is to provide perfect value to the customer through a perfect value creation process that has eliminated all unnecessary waste.

To accomplish this, lean thinking changes the focus of management from optimizing separate technologies and assets to optimizing the flow of the product or family of products through the entire value stream. Eliminating waste along the entire value stream, instead of at isolated points, creates processes that need less human effort, space, capital and time. This allows companies to make products and services at far lower costs and with fewer defects, compared with traditional business systems. Companies are able to respond to changing customer desires with great variety, high quality, low cost and very fast throughput times. Also, with the application of visual methods to control material flow and work-in-process, information management on the shop floor becomes much simpler and more accurate.

II. Procedure For The Implementation Of Mrp

The following procedures are followed while implementing Material Requirements Planning.

Demand for Products: the demand for end products stems from two main reasons. The first is known customers who have placed specific orders, such as those generated by sales

About-[1]Department of Mechanical Engineering, Federal University of Technology, Owerri, NIGERIA
(Tel: +234 803 668 3339 E-mail: ucheremy@yahoo.com)
Abou-t[2]Department of Mechanical Engineering, Faculty of Engineering, University of Port Harcourt, Choba, Rivers State, NIGERIA
(Tel: +234 806 234 0271 E-mail: james1bit@yahoo.com)

personnel, or from interdepartmental transactions. The second source is forecast demand.

Bill of Materials File: This is simply known as BOM file. It contains the complete product description, listing materials, parts, and components but also the sequence in which the product is created. The BOM file is often called the product structure file or product tree because it shows how a product is put together. It contains the information to identify each item and the quantity used per unit of the item of which it is a part.

Inventory Records File: Inventory records file under a computerized system can be quite lengthy. Each item in inventory is carried as a separate file and the range of details carried about an item is almost limitless. The MRP program accesses the status segment of the file according to specific time periods. These files are accessed as needed while running the program.

A. Conditions for implementation

Several requirements have to be met, in order to given an MRP implementation project a chance of success, among the conditions:

A. Availability of a computer based manufacturing system is a must. Although it is possible to obtain material requirements plan manually, it would be impossible to keep it up to date because of the highly dynamic nature of manufacturing environments.
B. A feasible master production schedule must be drawn up, or else the accumulated planned orders of components might mix with the resource restrictions and become infeasible.
C. The bills of material should be accurate. It is essential to update them promptly to reflect any engineering changes brought to the product. If a component part is omitted from the bill of material it will never be ordered by the system.
D. Inventory records should be a precise representation of reality, or else the netting process and the generation of planned orders become meaningless.
E. Lead times for all inventory items should be known and given to the MRP system.
F. Shop floor discipline is necessary to ensure that orders are processed in conformity with the established priorities. Otherwise, the lead times passed to MRP will not materialize.

B. Techniques for the implementation of MRP

MRP represents an innovation in the manufacturing environment. Thus, its effective implementation requires explicit management action. Steps need to be clearly identified and necessary measures be taken to ensure organizational responsiveness to the technique being implemented. Each organization poses a unique environment and that means that specific actions need to be taken with due regard to environment specifics.

We approach MRP as an organizational innovation and identify the necessary measure which management should adopt in implementing it. Motivational influences underlying MRP implementation include:

1. Recognition of business opportunity for the timely acquisition of MRP.
2. Recognition of technical opportunity for the timely acquisition of the technologies supporting MRP implementation.
3. Recognition of need for solving manufacturing and/or inventory problems using MRP. Given the above motivational factors one may readily identify what and how issues underlying MRP design and implementation.

What refers to a generic process model composed of steps and indicative levels of effort to implement each step.

How refers to management involvement with respect to the process.

C. Mrp Computer Program

The MRP program works as follows:

A. A list of end items needed by time periods is specified by the master production schedule.
B. A description of the materials and parts needed to make each item is specified in the bill of materials file.
C. The number of units of each item and material currently on hand and on order are contained in the inventory file.
D. The MRP program ―works" on the inventory file. In addition, it continuously refers to the bill of materials file to compute quantities of each item needed.
E. The number of units of each item required is then corrected for on hand amounts, and the net requirement is ―offset" to allow for the lead time needed to obtain the material.

D. Output Reports

Primary Reports: Primary reports are the main or normal reports used for the inventory and production control. These report consist of

1. Planned orders to be released at a future time.
2. Order release notices to execute the planned orders.
3. Changes in due dates of open orders due to rescheduling.
4. Cancellations or suspensions of open orders due to cancellation or suspension of orders on the master production schedule.
5. Inventory status data.

Secondary Reports: Additional reports, which are optional under the MRP system, fall into three main categories:

1. Planning reports to be used, for example, in forecasting inventory and specifying requirements over some future time horizon.

2. Performance reports for purposes of pointing out inactive items and determining the agreement between actual and programmed item lead times and between actual and programmed quantity usage and costs.

3. Exceptions reports that point out serious discrepancies, such as errors, out of range situations, late or overdue orders, excessive scrap, or nonexistent parts.

The Figure below shows an overall View of a Material Requirements Program and the Reports Generated by the Program.

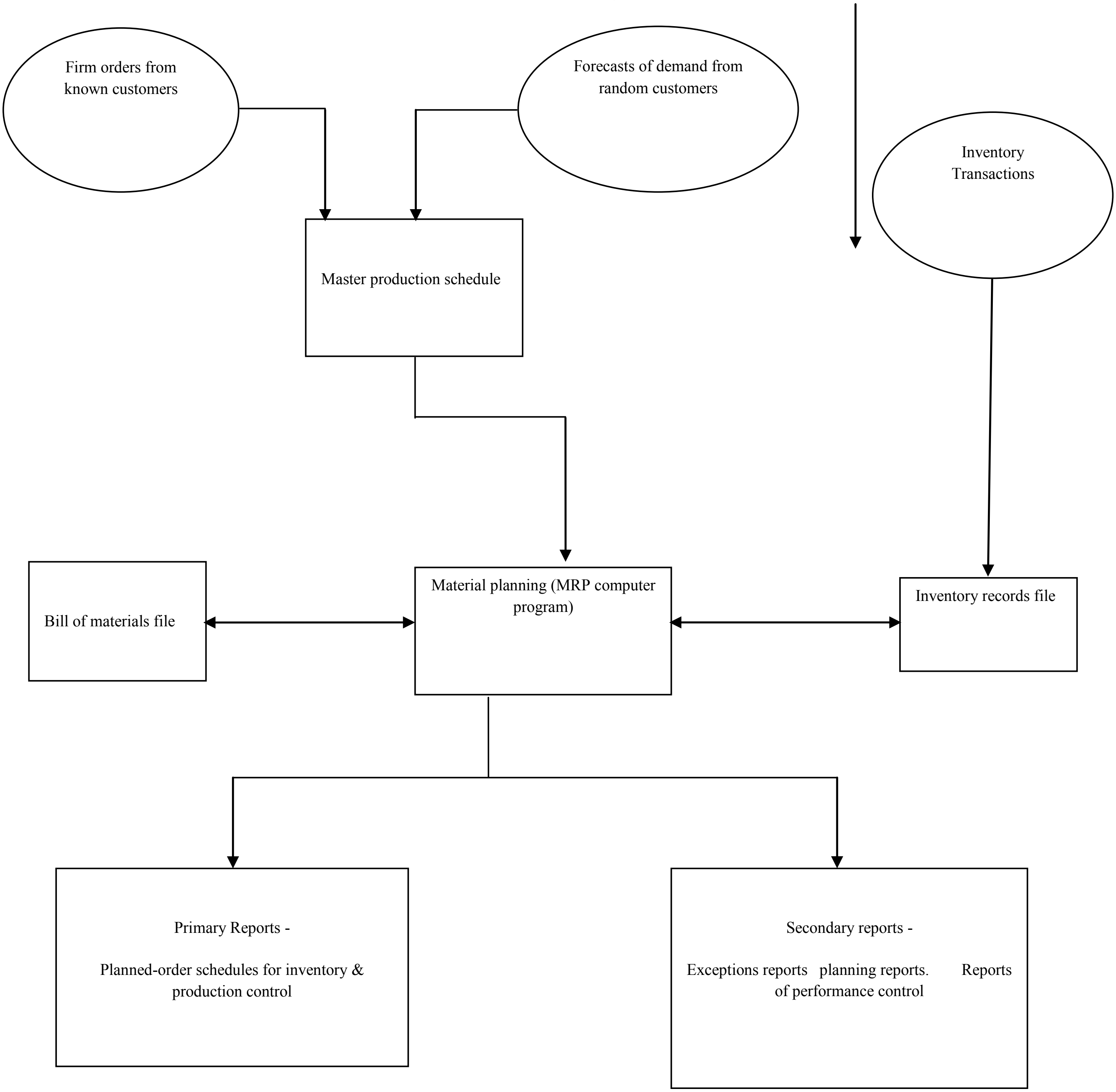

Firm orders from known customers
Forecasts of demand from random customers
Inventory Transactions
Master production schedule
Bill of materials file
Material planning (MRP computer program)
Inventory records file
Primary Reports -
Planned-order schedules for inventory & production control
Secondary reports -
Exceptions reports planning reports. Reports of performance control

Figure 1. Overall View of the Inputs to a Standard Material Requirements Program and the Reports Generated by the Program

E. MRP objectives

The main theme of MRP is –getting the right materials to the right place at the right time". Specific organizational objectives often associated with MRP design and implementation may be identified among three main dimensions, namely: inventory, priorities and capacity: Dimension: Objective specifics

Inventory

- Order the right part
- Order the right quantity
- Order at the right time

Priorities

- Order with the right due date
- Keep the due date valid

Capacity

- Plan for a complete load
- Plan for an accurate load
- Plan for an adequate time to view future load

III. LEAN MANUFACTURING

Lean manufacturing is a western adaptation of the Toyota Production System, developed by the Japanese carmaker and most famously studied (and the term ―Lean" coined) in The Machine That Changed the World (Womack, 1996). The Internet offers some useful resources on this topic, including BCG systems inc.(http//www.mmsonline.com), which state that Lean Manufacturing is a production method that calls for building products with as few steps and as little work-in-process inventory as possible. It relies on work centres or manufacturing cells that are capable of building multiple products, giving the company the flexibility to produce the exact mix and quantity of products required.

Taiichi Ohno, the engineer commonly credited with development of the Toyota Production System, and therefore Lean, identified seven types of waste: defective products, unnecessary finished products, unnecessary work in process, unnecessary processing, unnecessary movement (of people), unnecessary transportation (of products) and unnecessary delays. Lean focuses on eliminating these wastes from a manufacturing system. In particular, this work is interested in the second and third types – unnecessary finished goods and work in process. The Lean answer to these wastes is to link production at each step in the process with the subsequent process (or the consumer for finished goods). At Toyota, they use kanban (a Japanese word for ―shop sign") cards attached to each sub-assembly that are sent back to the producer each time one is used. The cards then become a signal to produce one more. As a result, the number of cards in the system controls the amount of work in process.

Liker (1997) describes a sequence of phases that a manufacturing facility must visit to become Lean: process stabilization, continuous flow, synchronous production, pull authorization, and level production. Such anecdotes are useful advice for managers and provide a general framework for becoming Lean, although they do not provide specific strategies for changing production control schemes.

Lean Manufacturing or Lean production, which is often known simple as ―Lean", is a production practice that considers the expenditure of resources for any goal other than the creation of value for the end customer to be wasteful, and thus a target for elimination.

According to Steinbrunner (2004), lean is centred on creating more value with less work. Lean manufacture is a generic process management philosophy derived mostly from the Toyota Production System (TPS) and identified as ―Lean" only in the 1990s. It is renowned for its focus on reduction of the original Toyota seven wastes to improve overall customer value, but there are varying perspectives on how this is best achieved. The steady growth of Toyota, from a small company to the world‗s largest automaker, has focused attention on how it has achieved this.

Lean manufacturing is a variation on the theme of efficiency based on optimizing flow; it is a present-day instance of the recurring theme in human history toward increasing efficiency, decreasing waste, and using empirical methods to decide what matters, rather than uncritically accepting pre-existing ideas. Lean manufacturing is often seen as a more refined version of earlier efficiency efforts, building upon the work of earlier leaders. A fundamental principle of lean manufacturing is demand-based flow manufacturing. In this type of production setting, inventory is only pulled through each production center when it is needed meet a customer‗s order. The benefits of this goal include: decreased cycle time, less inventory, increased productivity, increased capital equipment utilization.

The core of lean is founded on the concept of continuous product and process improvement and the elimination of non-value added activities. The value adding activities are simply only those things the customer is willing to pay for, everything else is waste, and should be eliminated, simplified, reduced, or integrated (Rizzardo, 2003).

Improving the flow of material through new ideal system layouts at the customer‗s required rate would reduce waste in material movement and inventory.

A. Steps to achieve lean systems

The following steps should be implemented to create the ideal **lean manufacturing** system:

1. Design a simple manufacturing system
2. Recognize that there is always room for improvement
3. Continuously improve the lean manufacturing system design

B. Basics for the design of a simple lean manufacturing system

A fundamental principle of lean manufacturing is demand-based flow manufacturing. In this type of production setting, inventory is only pulled through each production center when it is needed to meet a customer‗s order. The benefits of this goal include:

- decreased cycle time
- less inventory
- increased productivity
- increased capital equipment utilization

(a) There is always room for improvement

The core of lean is founded on the concept of continuous product and process improvement and the elimination of non-value added activities. ―The Value adding activities are simply only those things the customer is willing to pay for, everything else is waste, and should be eliminated, simplified, reduced, or integrated"(Rizzardo, 2003).

Improving the flow of material through new ideal system layouts at the customer's required rate would reduce waste in material movement and inventory.

(b) Continuously improve

A continuous improvement mindset is essential to reach a company's goals. The term "continuous improvement" means incremental improvement of products, processes, or services over time, with the goal of reducing waste to improve workplace functionality, customer service, or product performance (Suzaki, 1987).

C. Lean Goals

- The four goals of Lean manufacturing systems are to:
- Improve quality: To stay competitive in today's marketplace, a company must understand its customers' wants and needs and design processes to meet their expectations and requirements.
- Eliminate waste: Waste is any activity that consumes time, resources, or space but does not add any value to the product or service. There are seven types of waste:

1. Overproduction (occurs when production should have stopped)
2. Waiting (periods of inactivity)
3. Transport (unnecessary movement of materials)
4. Extra Processing (rework and reprocessing)
5. Inventory (excess inventory not directly required for current orders)
6. Motion (extra steps taken by employees because of inefficient layout)
7. Defects (do not conform to specifications or expectations)

- Reduce time: Reducing the time it takes to finish an activity from start to finish is one of the most effective ways to eliminate waste and lower costs.
- Reduce total costs: To minimize cost, a company must produce only to customer demand. Overproduction increases a company's inventory costs because of storage needs

IV. Discussion

MRP can be used to set priorities for the production of finished goods, in an environment where mixed mode is practised and in the job shop environment in order to develop a plan for common raw materials consumed. Uniform containers can be used to standardize lot sizes in production lines, for unique items consumed to signal the need to replenish materials and to simplify transport between the vendor and customer. Materials can then be pulled into the production lines as needed to support the required production rate of finished goods. Sharing material plans can lead to partnerships with vendors that not only reduce lot sizes and lead-times, but also result in reduced costs and less work-in-process at both vendor and customer locations. For the job shop environment, the planning and inventory tools of MRP can also be applied to set priorities for raw materials and manufactured products, in addition to developing plans for when and how much will be required.

Companies will continue to find ways to apply lean manufacturing concepts, if they should remain competitive, to simplify material planning, reduce waste and improve their operations. But it may not be feasible to apply pull methods to all of the company's product lines. When MRP planning and inventory tools are needed to support the job shop environment, and pull methods make sense to support the repetitive production lines, manufacturers will find that a blend of MRP push methods and lean manufacturing pull methods can provide the right material planning mix for their mixed mode environment. In order to have a successful implementation of MRP, the recommended steps are to be followed:

A computer based manufacturing system should be made available. It would be impossible to keep material requirements plan up to date because of the highly dynamic nature of manufacturing environments. Although it is possible to obtain material requirements plan manually, but it is time consuming and a daunting task.

A feasible master production schedule must be drawn up, or else the accumulated planned orders of components might fall into the resource restrictions and become infeasible.

The bills of material should be updated and accurate. It is essential to update BOM promptly to reflect any engineering changes brought to the product. If a component part is omitted from the bill of material it will never be ordered by the system.

Inventory records should be a precise representation of reality, or else the netting process and the generation of planned orders become meaningless.

Lead times for all inventory items should be known and given to the MRP system.

The last but not the least is maintaining Shop floor discipline. It is necessary to ensure that orders are processed in conformity with the established priorities. Otherwise, the lead times passed to MRP will not materialize.

V. Conclusion

MRP and lean are not only capable of co-existing, but they can also support one another, provided that the following concepts are understood and conditions exist:

Commitment to planning: First and foremost, there must be a commitment to planning. The "P" in MRP is for planning, yet its role is often overshadowed by the zeal to reduce waste. The importance of planning simply cannot be overlooked. Beyond better inventory control, planning enables you to have the right quality and quantity at the right location and time. Good material planning can help reduce the waste of downtime and reduce overtime. It also helps with overall product quality.

Communication with suppliers: While lean concepts reduce waste throughout every cycle of production, MRP can reduce waste in the supply chain through better relationships with suppliers. Planning enables better data and information that can be shared with vendors.

Dedication to data: While MRP systems can play an important role in synchronizing products, if changes occur, MRP can be slow to respond. This is usually a result of transactions not being entered in a timely manner. Effective product data management is critical to adapting traditional manufacturing systems to agile and lean manufacturing methods. However, it all begins with the data. By gaining an understanding about which bills of material and routing schemes are appropriate for given situations, you learn how they can be used to streamline operations, improve quality, reduce waste, minimize inventory and increase the use of manufacturing assets.

MRP is effective when people understand that the system cannot think for them. Too often, team members know that the information loaded into the system is useless, and they therefore have no faith in the resulting data that is intended to guide their ordering, systems, processes and operations - a classic case of garbage in, garbage out. However, if team members have confidence in the data, they will have confidence in the system.

Finally, when the principles are well integrated the following benefits will be obtained.

Improve quality: To stay competitive in today‘s marketplace, a company must understand its customers' wants and needs and design processes to meet their expectations and requirements.

Eliminate waste: Waste is any activity that consumes time, resources, or space but does not add any value to the product or service.

Reduce time: Reducing the time it takes to finish an activity from start to finish is one of the most effective ways to eliminate waste and lower costs.

Reduce total costs: To minimize cost, a company must produce only to the customer‘s specification and demand. Overproduction increases a company‘s inventory costs because of storage needs and inventory carrying cost.

VI. References

1) Agbu, O. (2007) The Iron and steel Industry and Nigeria‘s Industrialization:Exploring cooperation with Japan,institute of develping Economies,Chiba,Japan.
2) Auston, M.K. (1997) Lean manufacturing principles: A comprehensive Framework for improving production efficiency, University of california, los Angeles.
3) Black, J.T.and chen, J.C. (1994) *Decoupler-improved output of an apparel Assembly cell*, The journal of Applied manufacturing systems, winter, pp.47-58.
4) Edward, A.S., (1995) Inventory management and Production Planning and scheduling, John wiley and sons.
5) Gahagan, S.M. (2008) Simulation and optimization of production control for lean manufacturing Transition, unpublished dissertation submitted to the faculty of Graduate school, University of Maryland.
6) Jain,R.K. (2008) Production Technology, sixteenth Edition, Khanna publishers, 2-B,Nath market,Nai sarak,New Delhi.
7) James, H.G. (1997) American production and inventory control society production and inventory control Handbook, McGraw-Hill.
8) John, F.P. (1998) Master scheduling: A practical Guide to competitive management, John wiley and sons.
9) Liker, J.K. (1997) Becoming lean: inside stories of U.S. Manufacturers, Productivity press Portland, Oregon.
10) Mohommed, S.A.(2002) *African Iron and steel Industry*[online]. 10(8). Available from:http//globle.steel.com/[Accessed 22nd February 2010].
11) Moustakis, V.(2000) Material Requirements planning (MRP), Technical University of Crete.
12) Orlicky, J. (1976) Materials Requirements Planning, McGraw-Hill publisher.
13) Salem, O. and Zimmer, E. (2006) *Application of lean manufacturing principles to construction* [online].Available from:http//www.leanconstructionjournal.org/[Accessed 15th February 2010].
14) Steinbrunner, D. (2004) Modern *machine shop* [online].6(4).Available from:http//mmsonline .com/ [Accessed 2nd February 2010].
15) Waddell, B. (1984) International journal of production Research, vol.22,No. 2.pp. 193-233.
16) Womack, J.P. and Jones, D.T. (1996) Lean-thinking: Banish waste and create wealth in your company. New York.

A Study On Rough Clustering

Dr.K.Thangadurai[1] M.Uma[2] Dr.M.Punithavalli[3]

GJCST Computing Classification
H.3.3, I.5.3

***Abstract*-Clustering of data is an important data mining application. However, the data contained in today's databases is uncertain in nature. One of the problems with traditional partitioning clustering methods is that they partition the data into hard bound number of clusters. There have been recent advances in algorithms for clustering uncertain data, Rough set based Indiscernibility relation combined with indiscernibility graph, leads to knowledge discovery in an elegant way as it creates natural clusters in data. In this thesis, rough K-means clustering is studied and compared with the traditional K-means and weighted K-Means clustering methods for different data sets available in UCI data repository**

Keywords-Clusters, Boundary, Iteration, Attributes, Centroid.

I. Introduction

Clustering is a technique to group together a set of items having similar characteristic. There are two kinds of clusters to be discovered in web usage domain they are usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Clustering of pages will discover groups of pages having related content. This information is useful for internet search engines and web assistance providers. Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in collection of unlabeled data. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Here the simple graphical example for that

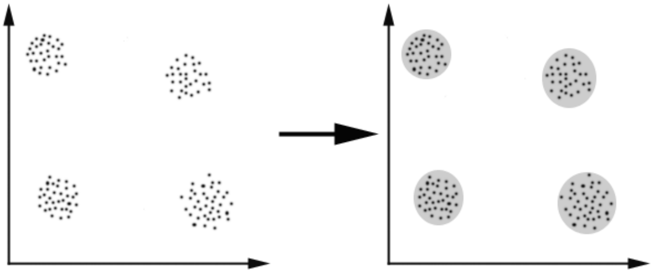


Figure 1: Cluster Analysis

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are ―close" according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures[1].

II. Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. There is no absolute best criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representative for homogeneous groups (data reduction), in finding natural clusters and describe their unknown properties (natural data types), in finding useful and suitable groupings (useful data class) or in finding unusual data objects (outlier detection).

A. The main requirements that a clustering algorithm should satisfy are

Scalability , dealing with different types of attributes, discovering clusters with arbitrary shape, minimal requirements for domain knowledge to determine input parameters, ability to deal with noise and outliers, insensitivity to order of input records, high dimensionality, interpretability and usability[2]

B. Numbers of problems with clustering are

Current clustering techniques do not address all the requirements adequately.

Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.

The effectiveness of the method depends on the definition of distance.

If and obvious distance measure doesn't exist we must define it, which is not always easy, especially in multi-dimensional spaces.

The result of the clustering algorithm can be interpreted in different ways

III. Clustering Algorithms

A large number of techniques have been proposed for forming clusters from distance matrices. The most important types are hierarchical techniques, optimization techniques and mixture models. We are going to discuss first two types here

About-[1] Head in Computer Science, Govt. Arts College (Men), Krishnagiri, TN, India(e-mail; ktramprasad04@yahoo.com)
About-[2] Research Scholar, Dravidian University, Kuppam, A.P., India
About-[3] Director, Department of Computer Science, SRCW, Coimbatore, TN, India

C. *Approaches to clustering*

1. Centroid approaches, 2.hierarchical approaches.
Centroid approaches: We guess the centroids or central point in each cluster, and assign points to the cluster of their nearest centroid.
Hierarchical approaches: We begin assuming that each point is a cluster by itself. We repeatedly merge nearby clusters, using some measure of how close two clusters are, or how good a cluster the resulting group would be.

D. *Hierarchical Clustering Algorithms*

A hierarchical algorithm yields a dendogram, representing the nested grouping of patterns and similarity levels at which groupings change. The dendogram can be broken at different levels to yield different clustering of the data. Most hierarchical clustering algoritms are variantas of the single-link, complete-link, and minimum-variance algorithms[3]. The single-link and complete-link algorithms are most popular. These two algorithms differ in the way of characterize the similarity between a parir of cluster.
In the single link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete link algorithm, the distance between two clusters is the minimum of all pair wise distance between patterns in the two clusters. The clusters obtained by the complete link algorithm are more compact then those obtained by the single link algorithm

IV. Partitional algorithms

A partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such a dendogram produced by a hierarchical technique. Parititional methods have advantages in applications involving large data sets for which the construction of a dendogram is computationally prohibitive. A problem accomanying the use os partitional algorithm is the choice of the number of desired output clusters. The partitional technique usually produce clusters by optimizing a citerion function defined either locally or globally.

A. *Clustering Techniques*

Let X be a data set, that is, X = {x_i=1……N}
Now let be the partition, $\mathfrak{R}$, of X into m sets, C_j , j=1…m. These sets are called clusters and need to satisfy the following conditions:
•$C_i \neq \emptyset$, i = 1... m
• $\bigcup_{i=1}^{m} C_i$ =X

•$C_i \cap C_j = \emptyset$, i ≠ j, i,j=1,….,m
It is important to say that the objects (vectors) contained in a cluster C_i are more similar to each other and less similar to the objects (vectors) contained in the other clusters. The intention in the clustering algorithms is to join (or separate) the most similar (or dissimilar) objects of a data set X, it is necessary to apply a function that can make a quantitative measure among vectors [8].
Partitional algorithm is typically run multipel times with different starting states, and the best configuration obtained from all of the runs issued as the output clustering.

B. *Types of partitional Algorithms*

- Squared Error Algorithms
- Graph-Theoretic Clustering
- Mixture-Resolving
- Mode-Seeking Algorithms

*K-Means Algorithm:*The K-means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described by Tou and Gonzalez.
1. Choose K initial cluster centre z1 (1), z2 (1)…zk (1).
2. At the k-th iterative step, distribute the samples {x} among the K clusters using the relation

$$x \in C_j(k) if \, \| x - z_j(k) \| < \| x - z_i(k) \|$$

For all i=1, 2…K; I ≠ j; where Cj (k) denotes the set of samples whose cluster centre is zj (k).
3. Compute the new cluster centre zj (k+1), j=1, 2…K such that the sum of the squared distances from all points in Cj (k) to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of Cj (k). Therefore, the new cluster centre is given by

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in C_j(k)} x$$

j=1, 2… K
Where N_j is the number of samples in C_j (k)
4. If z_j (k+1) =z_j (k) for j=1, 2…K then the algorithm has converged and the procedure is terminated.
5. Otherwise go to step 2

C. *Drawbacks of K-Means algorithm*

The final clusters do not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers.
We have to know how many clusters we will have at the first.

D. *Working Principle*

The K-Means algorithm working principles are clearly explained in the following algorithm steps.

Algorithm:

1) ***Initialize the number of clusters k.***
2) ***Randomly selecting the centroids in the given data set*** $(c_1 c_2 \dots c_k)$
3) ***Compute the distance between the centroids and objects using the Euclidean Distance equation.***
   a. d_{ij} = $||x_{i-}c_k||^2$
4) ***Update the centroids.***
5) ***Stop the process when the new centroids are nearer to old one.***
   Otherwise, go to step-3.

E. Weighted K-Means Algorithm

Weighted K-Means algorithm is one of the clustering algorithms, based on the K-Means algorithm calculating with weights. A natural extension of the K-Means problem allows us to include some more information, namely, a set of weights associated with the data points. These might represent a measure of importance, a frequency count, or some other information. This algorithm is same as normal K-Means algorithm just adding the weights. Weighted K-Means attempts to decompose a set of objects into a set of disjoint clusters, taking into consideration the fact that the numerical attributes of objects in the set often do not come from independent identical normal distribution.

The weighted k-means algorithm uses weight vector to decrease the affects of irrelevant attributes and reflect the semantic information of objects. Weighted K-Means algorithms are iterative and use hill-climbing to find an optimal solution (clustering), and thus usually converge to a local minimum.

In the Weighted K-Means algorithm, the weights can be classified into two types.

Dynamic Weights: In the dynamic weights, the weights are changed during the program.

Static Weights: In the static weights, the weights are not changed during the program.

The Weighted K-Means algorithm is used to clustering the objects. Using this algorithm we can also calculating the weights dynamically and clustering the data in the dataset.

Working Principle

The Weighted K-Means algorithm working procedure is same as the procedure for K –Means algorithm but the only weight is included in the weighted k means algorithm. The working procedure is given in the following algorithm steps.

Input: a set of n data points and the number of clusters (K)

Output: centroids of the K clusters

1. *Initialize the number of clusters k.*
2. *Randomly selecting the centroids* $(c_1 c_2 \dots c_k)$ *in the data set.*
3. *Choosing the Static weight* W *,which is range from 0 to 2.5 or (5.0)*
4. *Find the distance between the centroids using the Euclidean*
   Distance equation.

   $$d_{ij} = \left\| w_{.} * (x_{i\_} c_k) \right\|^{2}$$
5. *Update the centroids using this equation.*
6. *Stop the process when the new centroids are nearer to old one. Otherwise, go to step-4.*

F. Rough Set Clustering Algorithm

Rough sets were introduced by Zdzislaw Pawlak [6][7] to provide a systemic framework for studying imprecise and insufficient knowledge. Rough sets are used to develop efficient heuristics searching for relevant tolerance relations that allow extracting objects in data. An attribute-oriented rough sets technique reduces the computational complexity of learning processes and eliminates the unimportant or irrelevant attributes so that the knowledge discovery in database or in experimental data sets can be efficiently learned. Using rough sets, has been shown to be effective for revealing relationships within imprecise data, discovering dependencies among objects and attributes, evaluating the classificatory importance of attributes, removing data re-abundances, and generating decision rules [5]. Some classes, or categories, of objects in an information system cannot be distinguished in term of available attributes. They can only be roughly, or approximately, defined. The idea of rough sets is based on equivalence relations which partition a data set into equivalence classes, and consists of the approximation of a set by a pair of sets, called lower and upper approximations. The lower approximation of a given sets of attributes, can be classified as certainly belonging to the concept. The upper approximation of a set contains all objects that cannot be classified categorically as not belonging to the concept. A rough set also is defined as an approximation of a set, defined as a pair of sets: the upper and lower approximation of a set [7].

G. Rough K-Means Algorithm

Step 0: Initialization. Randomly assign each data object to exactly one lower approximation. By definition (Property 2) the data object also belongs to the upper approximation of the same cluster.

Step 1: Calculation of the new means. The means are calculated as follows:

$$m_k = \begin{cases} w_l \sum_{X_k \in C_k} \frac{X_n}{|C_k|} + w_B \sum_{X_k \in C_k^B} \frac{X_n}{|C_k^B|} & for\ C_k^B \neq \emptyset. \\ w_l \sum_{X_k \in C_k} \frac{X_n}{|C_k|} & Otherwise. \end{cases}$$

where the parameters wl and wb define the importance of the lower approximation and boundary area of the cluster. The expression |Ck| indicates the numbers of data objects in lower approximation of the cluster and |CBk | = |Ck −Ck| is the number of data objects in the boundary areas.

Step 2: Assign the data objects to the approximations. (i) For a given data object Xn
determine its closest mean mh:

$$d_{n,h}^{min} = d(X_n, m_k) = min_{k=1 \dots k}\, d(X_n, m_k)$$

Assign Xn to the upper approximation of the cluster h:Xn ∈ Ch.

(ii) Determine the means mt that are also close to Xn—they are not farther away from Xn than d(Xn,mh)where is a given threshold:

$$T = \{t : d(X_n, m_k) - d(X_n, m_h) \leq \varepsilon \cap h \neq k\}$$

If T= ∅ (Xn is also close to at least one other mean mt besides mh)
Then Xn ∈ Ct , ∀t ∈ T .
• Else Xn ∈ Ch.
Step 3: If the algorithms continue with Step 1.
Else STOP.

H. Experimental Results And Discussion

The experimental analysis is carried out in this chapter by considering three different data sets from UCI data depository and the algorithms are validated through XIE – BIEN index

I. Xie-Beni Validity Index

In this thesis, the Xie-Beni index has been chosen as the cluster validity measure because it has been shown to be able to detect the correct number of clusters in several experiments. Xie-Beni validity is the combination of two functions. The first calculates the compactness of data in the same cluster and the second computes the separateness of data in different clusters. Let S represent the overall validity index, π be the compactness and s be the separation of the rough k-partition of the data set. The Xie-Beni validity can now be expressed as:

$$\pi = \frac{\sum_{i=1}^{K}\sum_{j=1}^{n} {\mu_{ij}}^{2} \parallel x - z_i \parallel^{2}}{n}$$

Where
And s= $(d_{min})^2$
dmin is the minimum distance between cluster centres, given by
dmin= minij ||zi-zj||
Where n is the number of users, k is the number of clusters, and Zi is the cluster centre of cluster Ci, wl is taken as 0.7 for the elements that are placed in lower approximation, wu is taken 0.3 for the elements that are placed in Upper approximation, µij is taken as 0.3 for the elements that are placed in boundary region. µij be the membership value of the user in boundary region. Smaller values of π indicate that the clusters are more compact and larger values of s indicate the clusters are well separated. Thus a smaller S reflects that the clusters have greater separation from each other and are more compact. In this thesis, Xie-Beni validity index is used to validate the clusters obtained after applying the clustering algorithms

V. Conclusion

The K-Means, Weighted K-Means and Rough K-Means clustering algorithms have been studied and implemented. All the three algorithms are analyzed using the validity measure of Xie - Bien Index for three different UCI data sets. It is observed that Rough K-Means algorithm is performing well comparatively

VI. References

1) Agrawal R, Imielinski T and Swami A. ―Mining association rules between sets of items in large databases", In *Proc. 1993 Int. Conf. Management of Data (SIGMOD-93)*, 207-216. May 1993

2) Agrawal R, Mannila H, Srikant R, Toivonen H and Verkamo AI. ―Fast discovery of association rules.", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (Eds.) *Advances in Knowledge Discovery and Data Mining*, 307-328, 1996.
3) Bhattacharyya S, Pictet O, Zumbach G. ―Representational semantics for genetic programming based learning in high-frequency financial data.", *Genetic Programming 1998: Proc. 3rd Annual Conf.,* 11-16. Morgan Kaufmann, 1998.
4) Jiawei Han and Micheline Kamber, ―Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, USA, 2001.
5) Kusiak M, ―Rough set theory: *A Data Mining tool for semiconductor manufacturing",* IEEE Transactions on Electronics Packaging Manufacturing 24 (1) (2001) 44-50
6) Lingras P and West C, *"Interval set clustering of web users with rough K-means",* Journal of Intelligent Information Systems 23 (1) (2004) 5-16.
7) Lingras P, Yan R and M. Hogo, ―*Rough set based clustering: evolutionary, neural, and statistical approaches",* Proceedings of the First Indian International Conference on Artificial Intelligence (2003) 1074-1087.
8) Lingras, P. ―*Rough Set Clustering for Web Mining"*, Proceedings of 2002 IEEE International Conference on Fuzzy Systems. 2002.
9) .Milligan G.W and Cooper M.C., ―*An examination of procedures for determining the number of clusters in a data set",* Psychometrika, vol. 50, pp. 159-179, 1985.
10) Monmarche N. Slimane M, and Venturini G. Antclass, ―Discovery of cluster in numeric data by an hybridization of an ant colony with the k-means algorithm", Technical Report 213, Ecole d'Ingenieurs en Informatique pour l'Industrie (E3i), Universite de Tours, Jan. 1999.

Applying Software Metrics on Web Applications

Vikas Raheja[1] Rajan Saluja[2]

GJCST Computing Classification D.2.8

Abstract- **Web Applications Automates many daily business activities . User Interact with these web applications by the interface which these applications provides . Web applications are different from normal applications . The traditional software metrics can be applied to web applications . but some new metrics which are made only for web applications are important and increase the performance of web applications. In this paper traditional software metrics as well as some new web metrics are described . In new approach I have described the performance metric for web applications and security measures and navigability metric which are useful to improve the web applications . In the beginning I have given basics of measurements which are required for better understanding of this paper .**

Keywords-Web Metric ,Navigability Metric , Performance Metric , Security Metric

I. Introduction

Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules [2]. Software Metric is a term that embraces many activities , all of which involves some degree of measurement . Software Metrics provides a basis for improving the software process, increasing the accuracy of project estimates , enhancing project tracking , and improving software quality . There are many type of Software Metrics present out of which some are in the area of

1. Cost and Effort Estimation
2. Productivity Measure and Models
3. Data Collection
4. Quality Model and Measures
5. Reliability Models
6. Performance Evaluation and Models
7. Structural and Complexity metrics
8. Capability – Maturity Assessment

II. The Basic of Measurement

There are several theories of measurement, which will work like for e.g. Representational Theory of Measurement will work [2]. The Representational Theory of Measurement seeks to formulize our intuition about the way world works. that is , the data we obtain

as measures should represent the attributes of the entity. Our Intuition is the starting point for all measurements.

Empirical Relation : Given any two peoples x and y , we can observe that x is taller than y or y is taller than x therefore we say that ―Taller than is a empirical relation for height . where height is an attribute

Mapping: After finding the Empirical Relation one should go for mapping from Empirical Relation to Numerical Relation .

A is taller than B if and only if M(A) > M(B)

If we convert that type of relation to some mathematical form then such form is called mapping .

The stages for measurement are

- Identify attribute for some real world entities.
- Identify empirical relation for attributes.
- Identify numerical relations corresponding to each empirical relation.
- Define Mapping from real world entities to numbers.
- Check that numerical relations preserve and are preserved by empirical relation.

A. Direct and Indirect Measurement

Once we have a model of entities and attributes involved , we can define the measure in terms of them . Direct measurement of an attribute of an entity involves no other attribute or entity for example length of a physical object can be measured without reference to any other object or attribute . on the other hand , density of a physical object can be measured only in terms of mass and volume , we then use a model to show us that the relationship among the three is density = mass / volume . some direct measures in software engineering are length, duration of testing process , number of defects discovered ,time a programmer spends on the project. Indirect measurement is often useful in making visible the interactions between direct measurement [1] .

Example of Common Direct Measurement

Length

Width

Line of Code

Example of Common Indirect Measurement

Program Productivity: LOC produced / person months efforts

Module Defect Density: Number of Defects /module size

Requirements Stability: Number of initial requirement/ total number of requirements.

Test Effectiveness Ratio: Efforts spent fixing faults / total project effort

B. Measurement Scales and Scale types

There are five major type of scales .

- Nominal
- Ordinal
- Interval
- Ratio
- Absolute

About-[1] *Assistant Professor ,N.C. Institute of Computer Sciences , Israna , Panipat (Haryana)*

About-[2] *Assistant Professor ,N.C. Institute of Computer Sciences , Israna , Panipat (Haryana)*

C. *Classifying Software Measures*

Software measurement needs entities and attributes , we can divide our software to these three classes .
Processes : are collection of software-related activities.
Products : are any artifacts , deliverables or documents that result from a process activities
Resources : are entities required by the process activities.
With in each class of entities we distinguish internal and external attributes
Internal attributes : of a product , process or resources are those that can be measured purely in terms of the product , process or resources itself.
External attributes : of a product , process or resources are those that can be measured only with respect to how the product , process or resources relates to the environment .

Entities	Attributes	
Product	Internal	External
Specification	Size, Reuse, Modularity, Redundancy Functionality, Syntactic Correctness	Comprehensibility, Maintainability
Designs	Size , Reuse, Modularity, Coupling , Cohesiveness, Functionality	Reliability, Usability, Maintainability.
Code	Size, Reuse, Modularity, Coupling, Functionality, Algorithmic complexity, Control Flow Structure ness	Reliability, Usability, Maintainability,
Test data	Size , Coverage level	Quality
Processes		
Constructing Specification	Time, Effort, No of Requirements Changes ,	Quality , cost, Stability
Detailed Design	Time , Effort, No of Specification Faults Found	Cost, Cost Effectiveness
Testing	Time , Effort, No of Coding Faults Found	Cost, Cost Effectiveness, Stability
Resources		
Personal	Age, Price,	Productivity Experience, Intelligence
Teams	Size, Communication level, Structure ness	Productivity, Quality
Software	Price, size ,	Usability, Reliability
Hardware	Price, Speed , Memory Size,	Reliability,
Offices	Size, Temperature, Light	Comfort , Quality

Table 1

III. WEB METRICS

A. *Web Engineering Fundamentals*

Web Engineering is the implementation of engineering principals to obtain high quality web applications . Similar types of processes will be followed to make web applications as in traditional software's but with new ideas . now a day when the platform of programming has changed then it is difficult to develop the software only with traditional models . some changes in models needs to be required for the development of online applications. In the Previous years the web site consist of little more than a set of hypertext files that present information using text and limited graphics . as the time passed , HTML was augmented by development tools that enabled web engineers to provide computing capability along with information . As in traditional projects attributes are needed for software metrics either they are internal attributes or external attributes. Similarly attributes are needed by web metrics for the improvement of online projects or web applications . some of the attributes which are useful for web metric are
Network Intensiveness : A Web App resides on a network and must serve the needs of a diverse community of clients . Web Applications are network dependents [5].
Concurrency : A Large no of users may assess the Web Application at one time [5].
Unpredictable Load : At one time 1000 users can assess the web application or 10 users may assess the web application [5].

a. Performance : If a user wait for too long then , he or she may decide to go else where [5].

Availability : Web Application should be available for maximum time like 24/7/365 basis [5] .

b. Data Driven : The primary function of many web application is to present Hypermedia files as well as to display the graphics But web applications may also be able to assess the database [5].

Content Sensitive :The text present on the web sites should be of high quality . Because the contents always represent the quality of web sites [5].
Continuous Evolution : Web Applications evolves continuously. Some web applications may be updated after each hours ,some may be updated after each minutes .
Security : Web applications are on world network then there is need for securing the contents of web applications . strong security measures are to be taken for protecting the information and data of web applications
Meet the Business requirements web applications should solve the purpose of business for which they are made .
Various Types of Web Applications are

- Informational
- Downloads
- Customizable
- Interaction
- User Input
- Transaction Oriented
- Portal
- Database Access
- Data Warehousing

Table 2

B. *Planning For Web Engineering Projects*

In Table 2 the comparison of the traditional projects with small e- Projects and Major e-Projects has been carried out. Traditional Software Projects and Major e- Projects have substantial similarities . small e-projects have special characteristic which differs them form traditional projects . Even in case of small e-Projects planning must be occurred and risk must be considered , a schedule must be established and control must be defined so that confusion , frustration, and failure are avoided

	Traditional Projects	Small e-Projects	Major e-Projects
Requirement Gathering	Rigorous	Limited	Rigorous
Technical Specifications	Robust: Models spec	Descriptive Overview	Robust UML Models
Project Duration	Measured in month or years	Measured in days weeks or months	Measured in months and years
Testing & QA	Focused on achieving quality targets		
Risk Management	Explicit	Inherent	Explicit
Half Life Deliverables	18 months or longer	3 to 4 months	6 to 12 months
Release Process	Rigorous	Limited	Rigorous
Post Release – customer feedback	Requires Proactive efforts	Automatically obtained from user interaction	Obtained both automatically and solicited feedback

C. Project Management Issues for web applications

1) A Business must choose from one of the two web engineering issues (1) The web application is outsourced –The web Engineering is performed by some third party who has the expertise , talent and resources that may be lacking with in the business , (2) or the web application is developed in-house using web engineers that are employed by the business . A third alternative is there in which some work is carried out In-House and some work is outsourced [4] .

D. Our Approach Towards Web Metrics

Web Engineering uses metrics to improve the overall process for the development of web applications. These metrics provide the way how these web applications behaves and what is the quality of these online applications . Software Metrics provides a basis for improving the software process, Increasing the accuracy of project estimates , enhancing project tracking , and improving software quality . Web Metrics if properly characterized , achieve all these benefits also improve the usability , web Application performance , and user satisfaction [5].

The goal of web metrics is to provide better quality of web applications from technical and business point of view . Web Metric provides the measures of effort, time and complexity of web applications . Some of the measures of web applications are

E. Performance Metric

Performance is related to availability and concurrency of web applications . when end user require the service of web applications and web applications get fail such condition reduces the performance of Web applications . The cause of failure may be any thing either due to network failure or heavy load on servers Fig 1 shows an example of a typical web application architecture . in which web server take request from users and passes the request to database server through application server . and then result of database query will be shifted to client machine [8].

Single set of web server, application server and database server is giving the service to no of clients . with such type of architecture it is difficult to improve performance of web applications.

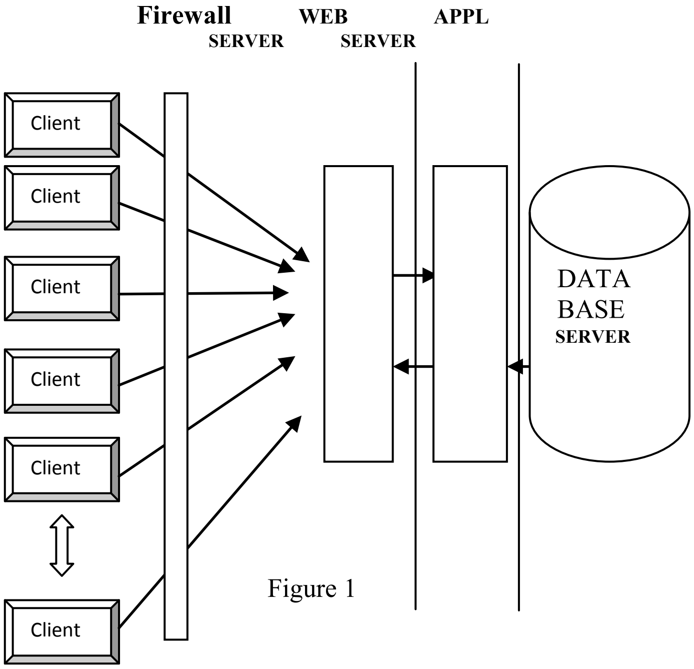


Figure 1

But if we improve such model and new model we make as shown in fig 2 in which two set of web server application server and database server give service to the clients then load on each server reduces our performance metric says that response time decreases if total no of servers increases To reduce the response time increase the no of servers.

Response Time α 1/ Total no of Servers

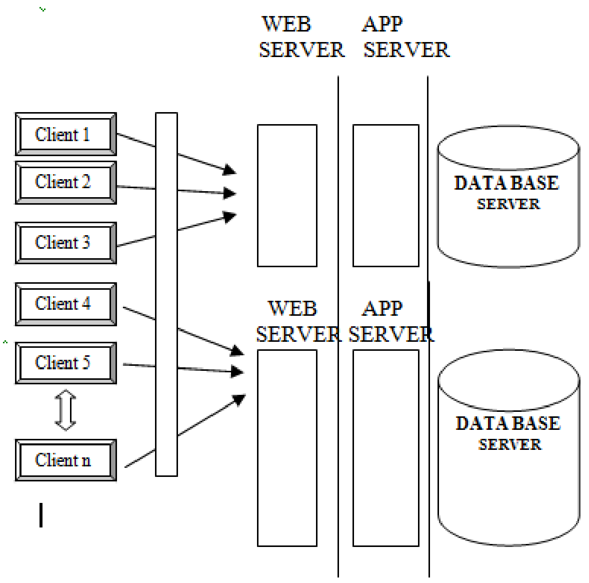


Figure 2

F. Security Metric

Web application are on world network then there is need for securing the contents of web applications [6]. Strong security measures should be taken to protect the information and data of web applications.

Inputs from user is the way through which security can be reduced . while coding the web applications appropriate checks should be implemented on user inputs to maintain the security of web applications .

e.g an input which is ready to take character type data should not take numeric data or any other special characters. Apply user ID and password on secure information .SQL Injection attacks which are done by hackers should be avoided by positive tainting techniques [2].

HTTP Cookies and server variables can be the cause for poor security . if user may not perform any action for some period of time then cookies should get expired and application should ask for relogin and password .Defensive programming reduces the attacks

IV. Measurement of Time and Efforts

A Few measures of efforts and time are given below Structuring efforts: Time to Structure Web Application Interlinking effort: Time to interlink pages to build the web applications, Interface Planning : Time taken to plan web application Interface, Interface Building : Time Taken to Implement interface for web applications Link-Testing effort: Time taken to test all links in web applications, Media Testing Time Taken to test all media in web Applications .

Total Effort:Structuring effort +Interlinking effort +Interface building +link-Testing effort + Media Testing efforts

(1) Page Authoring

Text Efforts : Time Taken to author or reuse text in page, Page linking efforts :Time Taken to author link in page, page structuring efforts :Time Taken to structure page.

Total page efforts: Text effort + page linking effort + page Structuring effort.

(2) Media Authoring

Media efforts: Time taken to author or reuse media files, Media digitizing: Time taken to digitize media Total media efforts: media effort + media digitizing effort [5] .

(3) Programming Authoring

Programming effort : Time taken to author HTML ,Java or related Language implementations Reuse
effort : Time taken to reuse/modify
existing programming .

(4) Navigability Measures

Navigability describe the ease with which user find the desired information . navigability measure is important for usability. A proper model of navigability reduces the access time .there are certain measures through which navigability can be increased e.g. hyperlinks depth ,hyperlinks breadths, topologies in connection with hyperlinks , some study have been done for the examination of hypertext topologies on usability [7] .breadth maximum approach all links are there on a single page or home page so that user can move to the desired page just by single click but this approach is better only for informational websites like Rediff home page .

Depth maximum approach all links are on different pages in a web application .depth is the no of clicks required to get the specific page from the home page . this approach is better where input is required from user by following the specific no of steps . web site navigability can be evaluated in three ways with user survey , with usage analysis , and with navigability measurements [7].

Mainly there are four hyper text topologies present (1) Linear Topology (2) Strictly Hierarchical (3) Mixed topology (Hierarchical Topology with cross –referential hyperlinks) (4) Non linear topology (a complete network based on a large no of cross referential links) .Previous study finds that navigability decreases in the order (1) Linear , (2) Strict (3) Mixed (4) Complex . we can divide the Mixed Topology into three sub categories (1) Mixed Hierarchical with link to Home Page (2) Bottom up approach (3) Mixed Hierarchical with link at the same level . In the first approach a link to home page is present from every page , In second approach a link to previous page is present from every page , and third approach is a link for every page at the same level is present .

V. References

1) E. Stroulia, M. El-Ramly, P. Iglinski, and P. Sorenson, ―User Interface Reverse Engineering in Support of Interface Migration to the Web," Automated Software Eng. J., vol. 10 no.3, pp. 271-301,2003.

2) Norman E . Fenton , ―Software Metircs ― Thomson Publications, Fifth Edition , 2005

3) Sreedevi Sampath,Lori Pollock —Apling Concept Analysis to User-Session-Based-Testing of Web Applications — IEEE Trans. on Software Engineering, Vol 33 , No. 10 pp. 643-657,Oct 2007
4) Powell, T.A. Web Site Engineering, Prentice Hall, 1998
5) Pressman Roger S, Software Engineering, McGraw Hill, 2005
6) Ying Zou, ,Qi Zhang, Xulin Zhao , —Improving the usability of e-commerce applications using business processes —, IEEE Transactions on Software Engineering ,Vol 33 , No 12 , pp. 837 – 853 ,Dec 2007
7) Yuming Zhou,Hareton Leung, —MNav : A Markov —,IEEE Trans. On Software Engineering ,Vol 33,No. 12, pp. 869-889 , Dec 2007.
William G.J. ,Alessandro Orso, Panagiotis ,” WASP
8) Protecting Web Applications Using Positive Tainting and Syntax Awareness —IEEE Trans. On Software Engineering ,Vol 34,No. 1, pp. 65-79 , Jan/Feb 2008.

Measuring Helpfulness of Personal Decision Aid Design Model

Siti Mahfuzah Sarif [1]

Norshuhada Shiratuddin [2]

GJCST Computing Classification
F.4.2, F.4.3

***Abstract*-The existence of countless computerized personal decision aids has triggered the interest to investigate which decision strategy and technique are ideal for a personal decision aid and how helpful is decision aid to non-expert users? Two categories of decision strategies have been reviewed; compensatory and non-compensatory, which results in fusing the two strategies in order to get the best of both worlds. Findings from the study of focus groups show that multi criteria decision method (MCDM) known as Pugh matrix and lexicographic have been identified as two most preferred techniques in solving personal decision problems. Both, the strategies and techniques, are incorporated in the development of a personal decision aid design model (PDADM). The proposed model is then validated through prototyping method in two different case studies (choosing development methodology in mobile computing course; and purchasing a mobile phone). In measuring the helpfulness of the prototypes, this study is looking at four dimensions; reliability, decision making effort, confidence, and decision process awareness. The findings show that the respondents from different decision situations perceived PDADM driven prototypes as helpful.**

Keywords-Computerized decision aid, decision strategy, multi criteria decision method, helpfulness

I. Introduction

Human commonly makes decisions of varying importance on daily basis, thus, making the idea of seeing personal decision making as a researchable matter seems odd. However, studies have proven that most humans are much poorer at decision making than they think. An understanding of what decision making involves, together with a few effective techniques, will help produce better decisions. Thus, explains the existence of decision support technology at different levels in various fields; for instance in management, engineering and medicine.

To date, the attentions given to the improvement of decision support at organization level has been enormous. On the contrary, the study in improving the performance of decision aid in personal decision making is still lacking and out of date (Jungermann, 1980; Wooler, 1982; Bronner & de Hoog, 1983; Alidrisi, 1987; Todd & Benbasat, 1991). The existence of countless computerized personal decision aids (either in the form of website, software or spreadsheet) these days, has triggered the interest to investigate the suitability and helpfulness of this technology to users, especially to the non-expert users.

II. Background of Study

Although most personal decisions made are minor in nature and in terms of its consequences, but still, being able to make an actual decision out of any situation is indeed essential (Rich,1999). Living in the 21st century, it is almost impossible not to associate anything with computer technology and this includes decision making. The evidence of human limitations in information processing is unquestionable, thus, the advantage of computerized decision aids can be a major benefit for decision maker.

A. Research Problem Statement

Decisions are part of human life. Decision majorly involves choices, and the hardest part is to make the right choice. It can be demanding to choose without being clear about what to choose and how to go about it, which later, may lead to being indecisive. Moreover, indecisiveness may cause failed actions and tendency of being controlled by others (McGuire, 2002; Arsham, 2004). This shows that, under appropriate circumstances, it is essential to apply decision aid in making decision.

Over decades, there are countless of studies on decision support technology that proposed the methods of improving the performance of such technology at organization level. However, in more recent years, the existence of computerized personal decision aids (more examples and reviews in section 3.2) are mushrooming and progressively getting attention from users; for example like ―hunch” (www.hunch.com) and ―Lt Simon Decide” (www.letsimondecide.com). This shows the relevance of study in issues related to computerized decision aids pertaining to personal decisions.

For more than five decades, most of research that have been carried out on decision process focuses either only on descriptive aspect (studying how decisions are being made) or normative aspect (studying how some ideally logical decider would make decisions). Decider in this context is referring to decision aid. Prescriptive research on decision processes, on how to help the decider progress from the descriptive to the normative has, however, been scarce (Brown, 2008). This is also has been mentioned earlier in (Bell et al., 1988).

About-[1] Department of Computer Science, College of Arts and Sciences, Universiti Utara Malaysia, and is currently pursuing her PhD degree. She specializes in software application development and multimedia design.
(email: ctmahfuzah@uum.edu.my)

About-[2] Professor and Applied Science Chair at the College of Arts and Sciences, Universiti Utara Malaysia. She obtained her PhD from University of Strathclyde, Glasgow, UK and has published more than 100 papers in journals and proceedings. She specializes in design research and application engineering.
(email: shuhada@uum.edu.my)

The term computerized decision aid refers to a very diverse set of tools based on a varying techniques and complexity. Generally, decision aids are designed with aims to help human choosing the best decision possible with the knowledge they have available. However, creating effective decision aids is more than meet the eyes (Power, 1998). Complex and structured mathematical techniques that correspond to the uncertainty of a decision situation have long held great theoretical appeal for helping decision makers make better decisions. Studies by Hayes and Akhavi (2008), Adam and Humphreys (2008), Zannier et al., (2007) and Law (1996) do not agree with the earlier statement. Hayes and Akhavi (2008) also affirmed that *"decision aids based on mathematically correct and sophisticated models do not actually improve the decision making performance. This is due to how the decision aids frame the problem in a way that does not fit human decision making approaches"*. Furthermore, although uncertainty can be tackled using complex mathematical tools, but more often than not, decision maker will not have the time to implement the structured mathematical strategies (McGuire, 2002; Arsham, 2004). These are further supported in Alidrisi (1987) and Adam and Humphreys (2008). All the researchers agreed that as far as personal decision making is concerned, complex and structured mathematical techniques are not preferred. Evidently, this indicates that a simple decision making model is a more needed solution when compared to the rigorous criteria weighing analysis.

All else being equal, decision makers prefer more accurate and less effortful choices. Since these desires are conflicting, thus selecting suitable strategy for the aid strategy can be a tricky task (Payne, 1993; Naude, 1997; Al-Shemmeri et al., 1997; Zanakis *et al.*, 1998). Then again, the appropriate use of decision strategies can contribute to effective decision making (Cosier & Dalton, 1986).

B. Research Objectives

With the nature of the problem in mind, this study aims to propose a personal decision aid design model that is perceived helpful. The following specific aims are outlined in means to support the general aim:

i. To identify the appropriate decision strategy and decision technique for personal decision making
ii. To incorporate identified decision strategy and technique in the development of the personal decision aid design model
iii. To validate the personal decision aid design model in different situations via prototyping method
iv. To measure the users' perceived helpfulness of the prototypes

III. Introduction to Decision Techniques

Apparently, a working knowledge of decision theory is needed before embarking into developing a decision aid design model. The design of the model includes two important expectations which are to accomplish a better decision and ensuring the helpfulness of the model via prototyping method.

Among the topics reviewed from the literatures include decision making, multi criteria decision making (MCDM) methods, computerized decision aids, related decision theories, and aspects of helpfulness of information systems in general and decision support in particular.

A. Decision Strategies and Techniques

Personal decision normally involves evaluation of many choices and making selection out of many. Generally, there are various strategies and techniques in making decision. This study focuses on decision making problems when the number of the criteria and alternatives is finite, and the alternatives are given explicitly. Problems of this type are called multi attribute decision making problems.

Compensatory and Non-compensatory Strategies

The decision strategies are commonly divided into two broad categories, non-compensatory and compensatory. Ullman (2002) defines non-compensatory strategies using the example of one well documented non-compensatory strategy; the lexicographic method.

As for compensatory strategies, Ullman (2002) defines it as strategy which allows decision makers to evaluate the alternatives by balancing the strong features of the alternatives with its weaker features. Example of methods that support compensatory strategy is decision matrix and utility theory methods.

Lexicographic method

In the lexicographic method, criteria are ranked in the order of their importance. The alternative with the best performance score on the most important criterion is chosen. If there are ties with respect to this criterion, the performance of the tied alternatives on the next most important criterion will be compared, and so on, till a unique alternative is found (Linkov *et al.*, 2004).

Maut

Multi-attribute utility theory (MAUT) is seen as an ideal approach for personal decision making by many previous researchers due to the nature of the decision problem. This is supported in a number of studies (Bronner & Hoog, 1983; Alidrisi, 1987; Işıklar & Büyüközkan, (2007); Adam & Humphreys, 2008). In a study, Adam and Humphreys (2008) described that, *"MAUT is simple enough to implement as compared to other model of decision making which requires a more rigorous criteria weighing analysis that is not necessarily needed for the role of decision making"*.

Pugh's Method

Pugh's method is known as the simplified MAUT which was first introduced by Pugh (1990) as the method for concept selection in engineering decision. In Pugh approach, all alternatives are compared to a datum alternative on each criterion. Alternatives are either better (+1), worse (-1), or each alternative is calculated as the number of occurrence of (+1) minus the occurrence of (-1). Emphasis was placed on using these comparisons to try to improve the weaknesses (i.e., the –1‘s) of an alternative without weakening any strength (i.e., +1‘s).

Weighted Decision Method

Weighted decision matrix involves mathematical reasoning in solving single or multi attribute decision problems. Two examples of weighted decision matrix are Weighted Sum Model (WSM) and Weighted Product Model (WPM). WSM is probably the most widely used approach, especially in single dimensional problems (Triantaphyllou, 2000). If there are *m* alternatives and *n* criteria then, the best alternative is the one that satisfies the following expression (Fishburn, 1967):

$$A^{*}_{WSMscore} = \max_{i} \sum_{j=1}^{n} a_{ij} w_{j}$$

, for i = 1, 2, 3 …m

WPM shares almost similar concept with WSM. The main difference is that instead of addition in the model there is multiplication. Each alternative is compared with the others by multiplying a number of ratios, one for each criterion. Each ration is raised to the power equivalent to the relative weight of the corresponding criterion. In general, in order to compare two alternatives A_K and A_L, the following product has to be calculated according to this expression (Bridgman, 1992; Miller & Star, 1969):

$$R(A_K \mid A_L) = \prod_{j=1}^{n} (a_{Kj} \mid a_{Lj})^{w_j}$$

,

where n is the number of criteria, aij is the actual value of i-th alternative in terms of j-th criterion, and wj is the weight of importance of the j-th criterion. If the term R (AK|AL) is greater than or equal to one, then it indicates that alternative AK is more desirable than alternative AL. The best alternative is the one that is better than or at least equal to all other alternatives.

the same (0) as the datum for a given criterion. The score for

Analytic Hierarchical Process

The Analytic Hierarchy Process (AHP) is a multi-criteria decision-making approach and was introduced by Saaty (1977 and 1994). The AHP has attracted the interest of many researchers mainly due to the careful mathematical properties of the method and the fact that the required input data are rather easy to obtain. The AHP is a decision support tool which can be used to solve complex decision problems. It uses a multi-level hierarchical structure of objectives, criteria, sub-criteria and alternatives.

Pros and Cons Analysis

Pros and Cons Analysis is a qualitative comparison method in which good things (pros) and bad things (cons) are identified about each alternative. Lists of the pros and cons, based on the input of subject matter experts, are compared one to another for each alternative. The alternative with the strongest pros and weakest cons is preferred. The decision documentation should include an exposition, which justifies why the preferred alternative‘s pros are more important and its cons are less consequential than those of the other alternatives. Pros and Cons Analysis is suitable for simple decisions with few alternatives and few discriminating criteria of approximately equal value. It requires no mathematical skill and can be implemented rapidly (Baker *et al.*, 2002).

B. Computerized Personal Decision Aids

A number of computerized decision aids have been identified. The aids come in varying mediums like website, spreadsheet, software and web application. All of the identified aids can be used to assist in personal decision making and also in other type of decision problems like financial and management problems. Table 3.1 summarizes eight computerized decision aids along with the reviews. The number of aids reviewed in this study is meant to be representative.

Table 3.1: Computerized decision aids

Decision aid	***Type***	***Method/ Technique***	***Description***	***Reviews***
1) Hunch (2009) (www.hunch.com)	Decision engine (web)	Collective intelligence decision making, machine learning & decision trees	• A decision community website • uses machine learning based on statistical inferences (the system gets smarter as more users use it) • uses question selection algorithm to a) find a question which will discriminate well among the remaining possible recommendation outcomes for user b) looks for a question which can help optimize and rank the remaining recommendation outcomes to present you with the ones you'll like the most	• the interactivity is intuitive but involves series of steps (answering questions) • Involves a lot of statistical analysis in the back end (very complex) • Does not involve defining importance of criteria (rank the criteria)
2) Let Simon Decide (2009) (www.letsimondecide.com)	Decision engine (web)	Collective intelligence decision, weighted decision analysis	• consists of three decision making tools: a. *My Scores*: for logical, fact based decision with multi-alternatives b. *My Life Match*: for big, life-changing decisions	• involves complex mathematical approach to decision-making • requires many steps

				c. *My Points of View*: for quick decision • combines user qualitative input with a weighted, mathematical formula (weighs alternatives against proprietary profile) • enables collective learning – share decision summary with others • provides action plan for every decision	although the process is intuitive
3)	Choose It! (1999) (chooseit.sitesell.com/)	Web application	Decision Matrix	• Online decision making tool that use decision matrix concept • can be used to make important business, financial, and personal life decisions	• does not acknowledge the distinct difference between subjective and objective factors
4)	Management For The Rest of Us (MFTROU.com) Decision Making Tool (n.d.) (www.mftrou.com/decision-making-tool.html)	Spreadsheet	Decision Matrix	• based on classic decision grid concept • in Excel spreadsheet format which contains: a. Overview of how to make decisions b. Decision Making Example c. Template for Making Your Own Decision	• crowded text in the visual presentation • Very formal presentation (in excel environment)
5)	Decision Oven (2008) (decisionoven.com/)	Software	Decision matrix with mathematical reasoning	• Off the shelf decision support software • can be used to support personal or business decisions	• acknowledge the difference between defining subjective criteria and objective criteria
6)	EduTools Decision Engine (2009) http://ocep.edutools.info/summative/index.jsp?pj=4	Web application	Weighted decision matrix	• use a rational decision making process	• Only focus on selecting a course management system, not for generic decision • User have to be familiar with the products and features that they wish to compare
7)	Career Decision Making Tool (CDMT) (n.d.) (http://cte.ed.gov/acrn/cdmt/tool.htm)	Instructor-led, classroom-based online tool	Guidelines and teaching/learning material	• It‘s a career decision making tool • It suggests the following decision cycles: a) Engaging b) Understanding c) Exploring d) Evaluating e) Acting f) Reflecting	• Only focus on career decision making, not for generic decision • To be implemented in teaching/learning environment
8)	Super Decisions (2004) http://www.superdecisions.com/	Software	Analytic Network Process	• It extends the Analytic Hierarchy Process (AHP) • Uses same fundamental prioritization process based on deriving priorities through judgments on pairs of elements or from direct measurements.	• Use complex decision analysis with rigorous mathematical reasoning • Solve for complex decision problem

C. Theories in Modeling Decision Aid Process

Decision theory is an attempt to explicate how human make decision, and in helping us understand the process of decision making. A grasp of the fundamentals of decision making is crucial to the effective design of the decision aid. Therefore, this study discusses a number of related theories that contribute to understanding multi criteria decision making. The related literature is summarized in Table 3.2.

Table 3.2: Literature survey of related decision theories

Decision Theories	References
Multi Attribute Utility Theory	Baker et al. (2001); Alidrisi (1987); Dyer et al. (1992); Keeney & Raiffa (1993); Collins et al. (2006)
Behavioral Decision Theory	Einhorn & Hogarth (1981); Westaby (2005)
Bounded Rationality Model	Bahl & Hunt (1984); March & Simon (1958); Newell & Simon (1972)
Implicit Favorite Model	Bahl & Hunt (1984); Soelberg (1967)
Dominance Theory	Easwaran (2007); Zsambok et al. (1992)
Satisficing Theory	Zsambok et al. (1992); Simon (1956)

IV. Research Methodology

This study employed design science approach to address the research questions posed earlier. The selection of a suitable approach is based on the nature of a research, phases involved and research outcomes. March and Smith (1995) described design science research as a process which aims to *"produce and apply knowledge of tasks or situations in order to create effective artifacts"* in order to enhance practice.

In general, process in design science research can be structured into three main phases include ―problem identification", ―solution design" and ―evaluation". Clearly, design science research consists of a series of steps but in practice they are not always executed in sequence; they often are performed iteratively. This study implemented the following steps, adapted from Offermann *et al.* (2009), and driven by design science research approach.

A. Problem Identification

The phase is divided into the following steps: ―identify problem", ―literature research" and ―expert interviews". It specifies a research question and verifies its practical relevance. As a result of this phase, the research questions are defined.

Identify Problem

The existence of countless computerized personal decision aids, these days, has triggered the interest to investigate the relevance and helpfulness of ICT assistance in personal decision making. Offermann et al. (2009) provides the support for the identification of research problem in this study, of which, they stated that researchable material ―may arise from a current business problem or opportunities offered by new technology".

Literature Search

In order to identify the research problem, literature search is used. As a summary, a number of decision strategies, decision techniques (MCDM methods), computerized personal decision aids, and decision making related theories were reviewed in this study. This results in strengthening the needs for a solution to propose a proper decision making model for personal decisions.

Expert Interview

Interviews with experts in the related field were conducted to identify relevancies of the addressed problems. Discussion with the experts involves brainstorming of idea, approval of idea and reviews on research material. Three experts have been referred to during this stage and also at certain stage of this study. The experts are professors and academics specializing in one of these fields: model-based systems and qualitative reasoning, quantitative analysis; and artificial intelligence.

B. Solution Design

In the second phase, the solution is designed and proposed. After identifying the research problems and evaluating its relevance, a solution is developed in the form of artifacts. Varying methods are used to come out with all the artifacts including content analysis, expert review, focus group study, participatory design, prototyping and elicitation work.

C. Evaluation

In this study, evaluation is achieved by the mean of case studies and laboratory experiments. The findings of this stage are further explained in Result section.

V. Development of Personal Decision Aid Design Model (PDADM)

This section describes the process in developing the PDADM. Prior to this, an appropriate decision strategies for personal decision making need to be identified, and followed by a selection of appropriate decision technique (i.e. MCDM method). Afterward, both will be incorporated in the development of the decision aid design model. The method used in developing PDADM involves content analysis, participatory design and expert review.

A. Decision Strategy Selection

From the literature search, two common decision strategy groups are studied; non-compensatory and compensatory. Findings indicate that non-compensatory strategies do not allow very good performance relative to one criterion to make up for poor performance on another. In other words, no matter how good an alternative is, if it fails on one evaluative criterion, it is eliminated from consideration.

As for compensatory strategies, they allow the decision makers to balance the good features of an alternative with its weaker features. Additionally, the compensatory strategies give greater accuracy in decision but the non-compensatory strategies take the least time to accomplish decision.

In responding to the earlier discussion, this study decided to combine the implementation of compensatory and non-compensatory strategies in order to obtain the ―best of both worlds". This is supported by Ullman (2002) in his work which stated that *"a method that gives the accuracy of the compensatory strategy with the effort of the non-compensatory strategy would add value to human decision making activities"*.

B. Decision Technique Selection

In light of the numerous decision techniques available to decision makers, study of focus groups is used in order to get some understanding of which kind of techniques that is more preferred by the (non-expert) decision maker. This study also decided that introducing more than one would enhance focus groups abilities to understand that there is not a single right way to resolve a decision.

There are five techniques that were introduced to the focus group of 51 (non-expert) participants of varying demographic background; weighted sum method (WSM), Pugh matrix (PUG), Analytic Hierarchy Process (AHP), pro and cons analysis (PCA), and lexicographic (LEX). All methods involve defining criteria on which to compare a set of alternatives. The group was encouraged to solve the same decision scenario (choosing a laptop from 4 different brands) using each or at least three of the techniques mentioned above one at a time. This study did not make it compulsory for them to use all the techniques, because of varying rate of understanding of the techniques after first time being introduced to them. Hence, unutilized techniques show respondents‘ difficulty to understand and to get familiar with it.

After establishing the focus group previous experience with each decision technique, the group was asked which technique helped the most and which they had more confidence in. Next, the group was asked which tool they think is ―least prone to bias”.

The results from the survey are summarized for each question. The first two questions concerned (i) which technique that they think helped the most if they were to use it in real decision and (ii) which technique they had the most confidence in. As shown in Table 5.1, technique PUG and LEX scored among the highest number of respondents for both questions.

Table 5.1: Helpful and Confidence

	WSM	PUG	AHP	PCA	LEX
Helpful	21	39	3	19	43
More confidence in	14	31	3	15	45

The next question asked the group which technique they felt was least prone to bias (that is, is the most difficult to manipulate to achieve preconceived results). These results are shown in Table 5.2.

Table 5.2: Bias

	WSM	*PUG*	*AHP*	*PCA*	*LEX*
Least prone to bias	34	41	2	18	22

Interestingly, even though majority of the participants had more confidence in LEX, the score changes when it comes to biasness of the technique. More than half of them felt that PUG was less prone followed by second the highest scored technique; the WSM. Nevertheless, the participants noted that it would take even more time and effort to achieve decision with the PUG and WSM. It is noted that AHP scores the lowest response for all three questions, which is due to refusal of most respondents to utilize it. Evidently, from this focus group study, PUG and LEX are selected as the potential techniques to be incorporated in the design of proposed personal decision aid design model.

PUG or Pugh matrix is originally a concept selection method used by engineers for design decision (Pugh, 1990). Since it was introduced, there have been many different modified versions of Pugh matrix analysis in various examples of its applications. In line with this, a participatory design study was conducted to learn which implementation of the Pugh matrix is preferred and suitable with the non-expert decision is making style. There are five versions (see Appendix) of Pugh matrix approach (including the original) used in this participatory design study. A total of 66 participants of varying demographic background were involved in this study.

Firstly, the participants were briefly explained about the different implementations of the Pugh matrix method. Then, they were asked to solve a designated decision problem (choosing a laptop from four different brands) using all four versions; one at a time. Later, the participants were asked ten questions (refer Table 5.3) based on their experience using the different implementation of Pugh matrix and also three additional demographic questions on gender, IT skill and age.

Table 5.3: Questions asked in the participatory design study

No.	Question
Q1	Are you familiar with the use of Pugh matrix?
Q2	Do you find it difficult to choose the first reference?
Q3	Do you prefer to weigh or not to weigh the criteria?
Q4	Do you prefer to use percentage (%) or scaled values (e.g. 1 to 5) as weight?
Q5	Do you prefer to use comparative symbols (+, -, S) or scaled values (e.g. 1 to 5) to rate the alternatives?
Q6	Which version of Pugh matrix do you think is most helpful?
Q7	Which version of Pugh matrix you had more confidence in?
Q8	In your opinion, which version is least prone to bias?
Q9	Would you use either of these Pugh matrix approach in your real life decision?
Q10	Would it be easier if Pugh matrix process is automated (i.e. in a computerized format)?

All the responses from participants were recorded and summarized in the following tables (Table 5.4 to 5.12). The first question dealt with the previous experience of the participants with Pugh matrix method. As shown in Table 5.4, majority of the participants had not used the Pugh approach before this study.

Table 5.4: Familiar with Pugh matrix

	Yes	*No*	*NA**
Familiar?	9	57	0

**=No answer*

The next question asked about participants experience during the study when they were required to choose their own reference for comparative analysis in Pugh matrix take place. As shown in Table 5.5, more than half of the participants claimed that it is not a problem for them to perform that task. But the number of participants who claimed the opposite was not far behind.

Table 5.5: Difficulty to choose first reference

	Yes	*No*	*NA*
Difficult?	24	42	0

The third and fourth questions asked about participants experience with the use of weight in defining the importance of each of the evaluative criteria. As shown in Table 5.6, majority of the participants preferred to weigh their criteria during the process. From this majority group, 35 of them preferred weighing the criteria using scaled values than using percentage (Table 5.7). This number represented more than half of the participants.

Table 5.6: Weighing criteria

	Yes	*No*	*NA*
Weighing criteria	42	21	3

Table 5.7: Use percentage or scaled values for weighing

	Percentage	*Scaled Values*	*NA*
Preferred weighing criteria	26	35	5

The fifth question asked the participants if they prefer to use symbols; + for better, - for worse and + for equal); or scaled value to perform the comparative analysis of alternatives against the reference on each criterion. Majority agreed that the use of symbols is more convenience for the comparative analysis.

Table 5.8: Use symbols or scaled values

	Symbols	*Scaled Values*	*NA*
Preferred evaluation styles	52	12	2

The next two questions (question 6 and 7) dealt with participants experience after using the Pugh approach to solve the decision problem. As shown in Table 5.9, the obviously dominant choice for both questions is the original version. The participants, as a whole, not only felt like the original version helped the most in assisting them with decision problem, but they had more confidence in it.

Table 5.9: Helpful and confidence

	Original	*MV1*	*MV2*	*MV3*	*MV4*	*NA*
Helpful	22	11	13	7	8	5
More confidence in	21	10	14	8	10	3

MV=modified version

Even though majority has more confidence in the original version, but when asked about which version they think is least prone to bias, the majority score shows contrasting response. One third of the participants agreed MV2 (modified version #2) is the one least prone to bias.

Table 5.10: Bias

	Original	*MV1*	*MV2*	*MV3*	*MV4*	*NA*
Least prone to bias	15	11	22	10	4	4

Concerning the use of Pugh approach in real decision situation, 49 of 66 indicated that they will consider using this approach, 16 indicated that they would not, and one did not respond to this question (refer Table 5.11).

Table 5.11: Will use Pugh matrix in real situation

	Yes	*No*	*NA*
Will use Pugh approach in real situation?	49	16	1

Lastly, when asked whether the participants think that by automating the process of Pugh matrix (in computerized format) will make it easier to use this approach, majority of them answered yes. From 12 of the remaining participants who answered no, 7 of them appeared to claim themselves as having very less IT skill.

Table 5.12: Automate Pugh matrix

	Yes	*No*	*NA*
Automating Pugh approach makes it easier?	54	12	0

(5)

C. Incorporating the Decision Strategy and Decision Technique in PDADM

The results; decision strategies and techniques, obtained from previous focus group study are incorporated in the development of personal decision aid design model. The model comprises of the flow of the decision process and the relationship between input and outcome of each step of the process. Figure 5.1 illustrates the previous statement clearer.

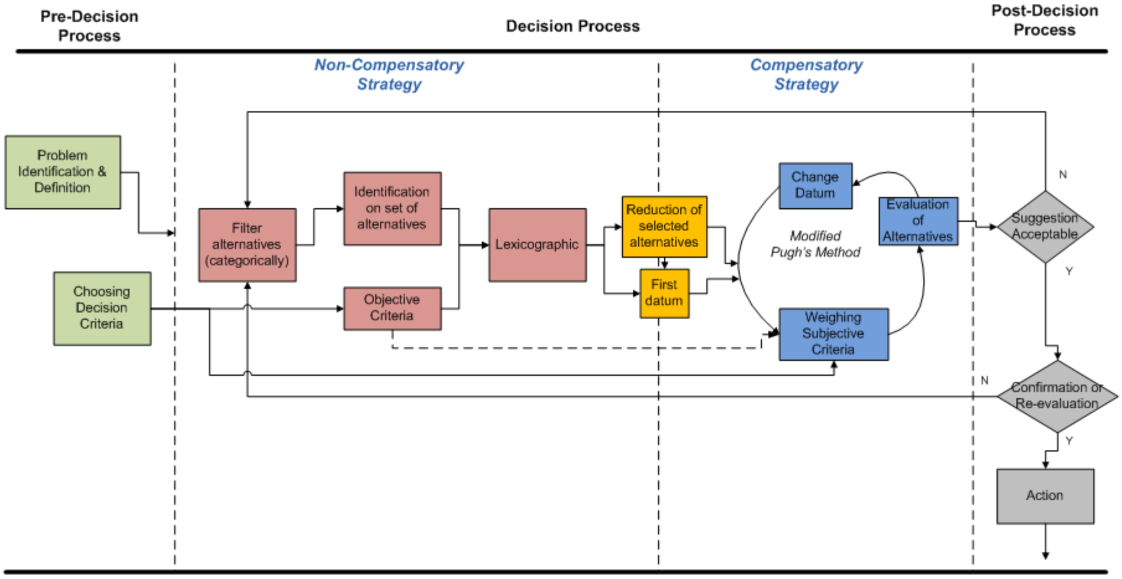

Pre-Decision Process
Decision Process
Post-Decision Process
Non-Compensatory Strategy
Compensatory Strategy
Problem Identification & Definition
Choosing Decision Criteria
Filter alternatives (categorically)
Identification on set of alternatives
Objective Criteria
Lexicographic
Reduction of selected alternatives
First datum
Change Datum
Evaluation of Alternatives
Modified Pugh's Method
Weighing Subjective Criteria
Suggestion Acceptable
N
Y
Confirmation or Re-evaluation
N
Y
Action

Figure 5.1: Personal Decision Aid Design Model (PDADM)

VI. Implementing PDADM in Different Situations

The proposed PDADM is validated through development of two prototypes in two different case studies; choosing development methodology in mobile computing course; and purchasing a mobile phone. These case studies involved two very different decision situations which were intended to showcase the flexibility and functionality of the proposed model.

A. Case study 1: Choosing a Development Methodology in Mobile Programming Course

Over the last decade, mobile computing has received significant interest in the academic and industrial research community. As a result, demands from the industry for graduates of mobile computing course are rising (Gillespie, 2007).

The graduates who are entering the mobile development world are expected to put up with the challenges imposed by the mobile environment. Heyes (2002) reported that mobile developers face twice as much as challenges than developing traditional system application due to the specific demand and technical constraints of mobile environment. In addition to that, inadequate research in assisting developers with the mobile development issues is also highlighted in the GI Dagstuhl Research Seminar in 2007 (König-Ries, 2009). Within this perspective, it is believed that selecting a suitable development methodology is the key to these issues. The use of a methodology is important, as a project can be structured into small, well-defined activities where the sequence and interaction of these activities can be specified (Avison & Fitzgerald, 1990). Hence, students should be exposed to the importance of adopting a suitable methodology for a mobile development project.

development project is another challenge in itself (Bertini *et al.*, 2006; Heikkinen & Still, 2005; Atkinson & Olla, 2004; Heyes, 2002; Afonso *et al.*, 1998). Less experienced developers will find the task even more challenging, thus, this study seeks to propose a solution by implementing the proposed PDADM via a development of prototype named as m^d-Matrix (as in mobile development methodology matrix).

Features and Screenshots of m^d-Matrix

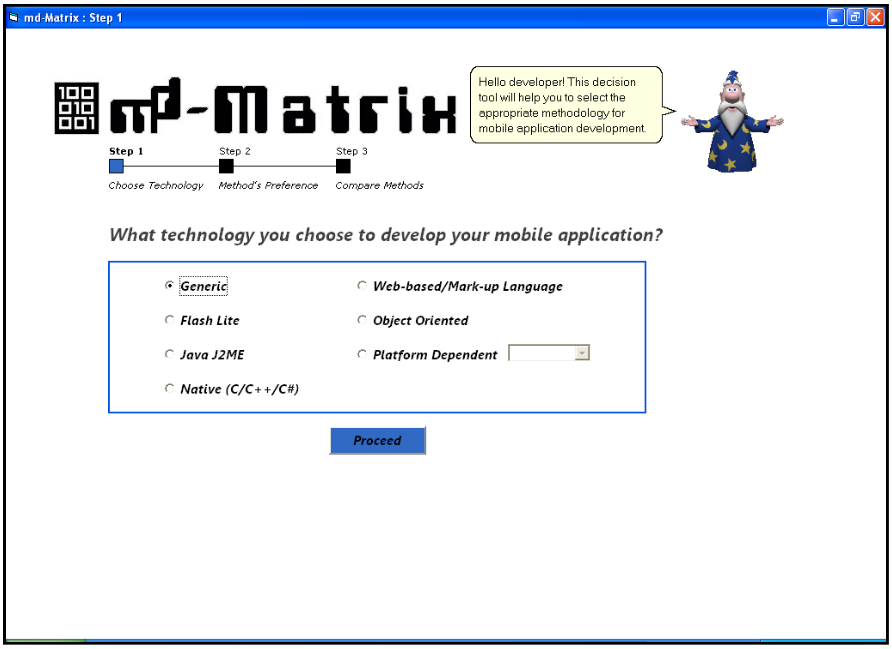
This decision-making tool is mainly aimed at assisting developers (especially the novice) in choosing the most appropriate development methodology for mobile development project. The numbers of available development methodologies in md-Matrix are meant to be representative; only for the purpose of demonstrating the decision process that occur in selecting a mobile development methodology. The prototype of md-Matrix features the following (see Table 6.1):

Table 6.1: Features of m^d-Matrix

m^d-Matrix	
Alternatives filter	Mobile application technologies: Generic J2ME* Flash Lite* Native Web based Object Oriented Platform dependent
Criteria	12 objective 12 subjective
Alternatives	Flash Lite (4 methodologies) J2ME (4 methodologies)
Feedback	Pop-up window On screen text Interface agent

* *enabled in this prototype*

The first step of m^d-Matrix enables user to filter the available methodologies based on preferred technology for development of a mobile application (Figure 6.1). As it proceeds with the second step (Figure 6.2), users will make their selection of narrative criteria to further filter the options (methodologies) following the non-compensatory strategy (lexicographic process). The three highest scored methods (see Figure 6.3) which pass most of the selected criteria will be ranked accordingly and the one in the highest rank will be set as the first reference (datum). Next, the three identified methods from previous step will be compared to each other following the compensatory strategy (modified Pugh‗s method) based on preferred subjective criteria (Figure 6.4). The steps can be iterated in maximum 3 cycles where in each round the reference will be changed until each methodology will be a reference once. The dominance methodology from the 3 rounds will be suggested as the best selection. The following are screenshots of m^d-Matrix:

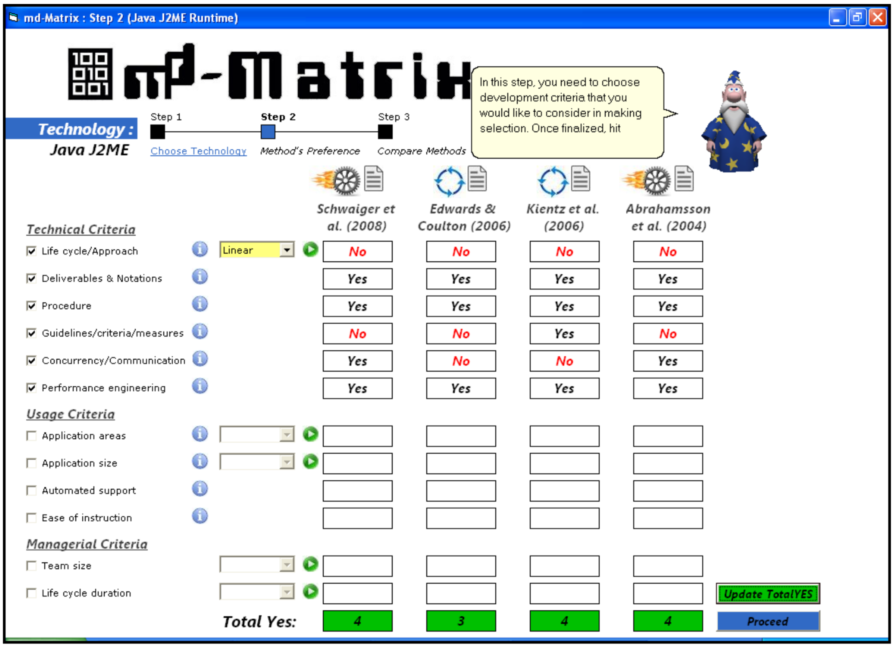


Hello developer! This decision tool will help you to select the appropriate methodology for mobile application development.

What technology you choose to develop your mobile application?

● Generic
○ Flash Lite
○ Java J2ME
○ Native (C/C++/C#)
○ Web-based/Mark-up Language
○ Object Oriented
○ Platform Dependent

Figure 6.1: Alternatives filtered categorically

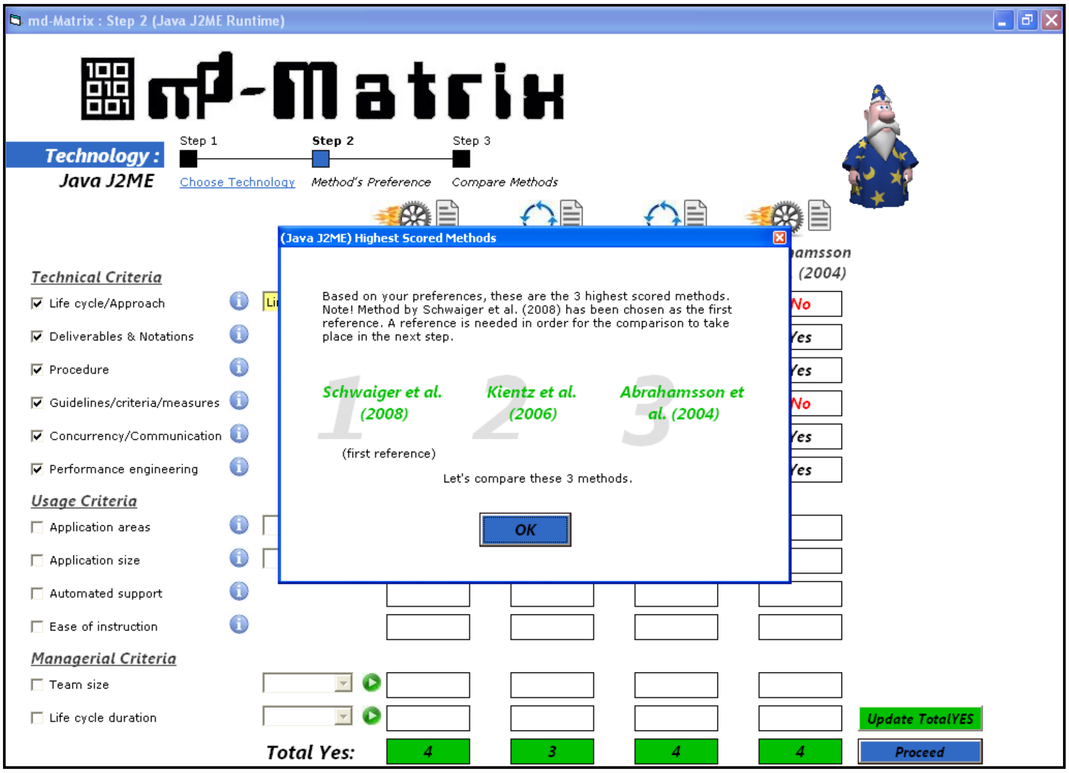


Technology : Java J2ME

In this step, you need to choose development criteria that you would like to consider in making selection. Once finalized, hit

		Schwaiger et al. (2008)	Edwards & Coulton (2006)	Kientz et al. (2006)	Abrahamsson et al. (2004)
Technical Criteria					
☑ Life cycle/Approach	Linear	No	No	No	No
☑ Deliverables & Notations		Yes	Yes	Yes	Yes
☑ Procedure		Yes	Yes	Yes	Yes
☑ Guidelines/criteria/measures		No	No	Yes	No
☑ Concurrency/Communication		Yes	No	No	Yes
☑ Performance engineering		Yes	Yes	Yes	Yes
Usage Criteria					
☐ Application areas					
☐ Application size					
☐ Automated support					
☐ Ease of instruction					
Managerial Criteria					
☐ Team size					
☐ Life cycle duration					
	Total Yes:	4	3	4	4

Figure 6.2: The 12 objective criteria used in non-compensatory (lexicographic) process



md-Matrix : Step 2 (Java J2ME Runtime)

Technology : Java J2ME

Step 1 Choose Technology — Step 2 Method's Preference — Step 3 Compare Methods

(Java J2ME) Highest Scored Methods

Based on your preferences, these are the 3 highest scored methods. Note! Method by Schwaiger et al. (2008) has been chosen as the first reference. A reference is needed in order for the comparison to take place in the next step.

1 Schwaiger et al. (2008) (first reference)

2 Kientz et al. (2006)

3 Abrahamsson et al. (2004)

Let's compare these 3 methods.

OK

Technical Criteria

☑ Life cycle/Approach
☑ Deliverables & Notations
☑ Procedure
☑ Guidelines/criteria/measures
☑ Concurrency/Communication
☑ Performance engineering

Usage Criteria

☐ Application areas
☐ Application size
☐ Automated support
☐ Ease of instruction

Managerial Criteria

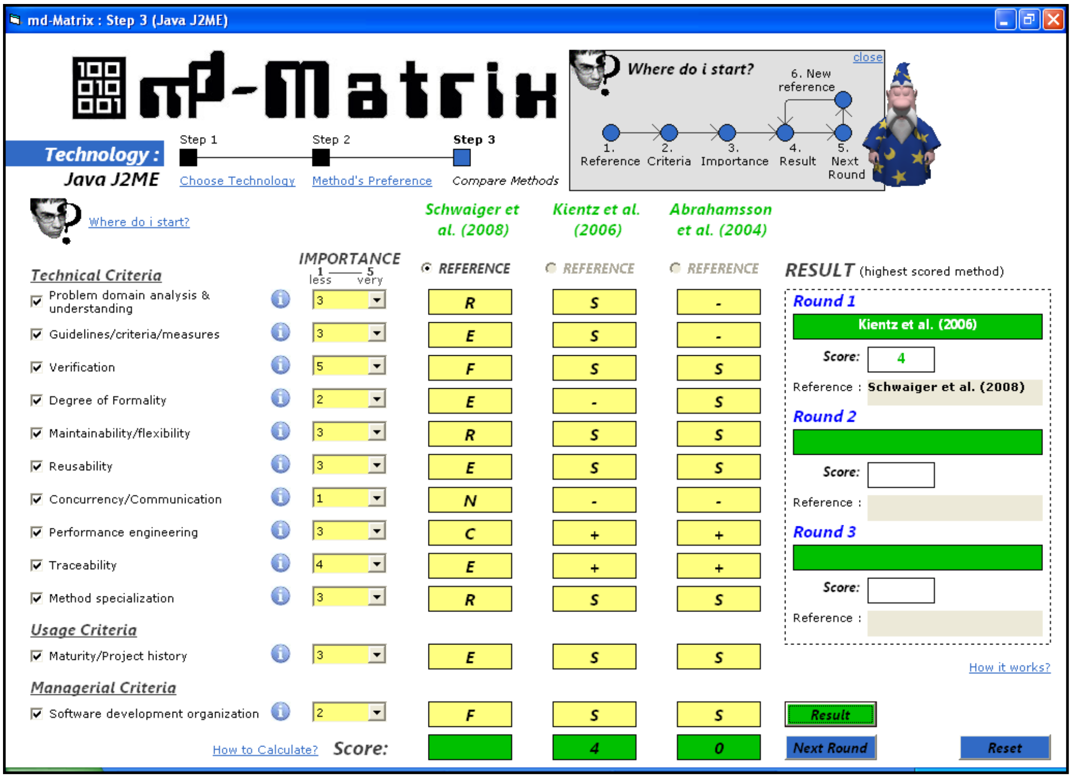
☐ Team size
☐ Life cycle duration

Update TotalYES

Total Yes: 4 | 3 | 4 | 4

Proceed

Figure 6.3: Result obtained in non-compensatory process



md-Matrix : Step 3 (Java J2ME)

Technology : Java J2ME

Step 1 Choose Technology — Step 2 Method's Preference — Step 3 Compare Methods

Where do i start? 1. Reference 2. Criteria 3. Importance 4. Result 5. Next Round 6. New reference close

	IMPORTANCE (1 less — 5 very)	Schwaiger et al. (2008) ● REFERENCE	Kientz et al. (2006) ○ REFERENCE	Abrahamsson et al. (2004) ○ REFERENCE
Technical Criteria				
☑ Problem domain analysis & understanding	3	R	S	-
☑ Guidelines/criteria/measures	3	E	S	-
☑ Verification	5	F	S	S
☑ Degree of Formality	2	E	-	S
☑ Maintainability/flexibility	3	R	S	S
☑ Reusability	3	E	S	S
☑ Concurrency/Communication	1	N	-	-
☑ Performance engineering	3	C	+	+
☑ Traceability	4	E	+	+
☑ Method specialization	3	R	S	S
Usage Criteria				
☑ Maturity/Project history	3	E	S	S
Managerial Criteria				
☑ Software development organization	2	F	S	S
How to Calculate? Score:			4	0

RESULT (highest scored method)

Round 1: Kientz et al. (2006)
Score: 4
Reference : Schwaiger et al. (2008)

Round 2: Score: Reference :

Round 3: Score: Reference :

How it works?

Result | Next Round | Reset

Figure 6.4: The 12 subjective criteria used in compensatory process

m^d-Matrix as a Learning Tool

Along providing solution to the selection of development methodology, md-Matrix also can be utilized as an educational tool either in academic or industry. Learning institutions can utilize it for teaching purposes to educate students on the need to have a well-structured process of developing mobile applications. As for the industry, this tool can be used as one of the materials for training of new interns and apprentice developers.

B. Case study 2: Choosing a Mobile Phone

Consumers are faced with purchase decisions mostly every time when a purchase is required. But not all decisions are treated the same. Some decisions are more complex than others and thus require more effort by the consumer. Other decisions are fairly frequent and require little effort.

Consumers will not simply go to a store or online catalog and spend their money in a rush. Purchasing takes place usually as a result of series of decision making steps. The implication of buying behavior shows the need for a reliable decision making tool to assist consumers in making a less-regretful and effective decision (Häubl & Trifts, 2000; Chris, 2008).

It is also important for the consumers to be able to decide on the purchasing item with confidence and ease. Thus, a comprehensive and undemanding decision aid is much needed in the process. Another important aspect is the use of decision aid in raising awareness about the consequences of actually choosing the item and purchases it. This could be obtained by organizing data with the purpose of presenting or displaying it to the decision maker (consumer) in a much clearer way than simply making a list of the alternatives. Within this perspective, the proposed PDADM is implemented in assisting consumers to make purchasing decision via the use of the prototype known as e^{p}-Matrix (as in electronic purchasing matrix).

Features and Screenshots of e^{p}-Matrix

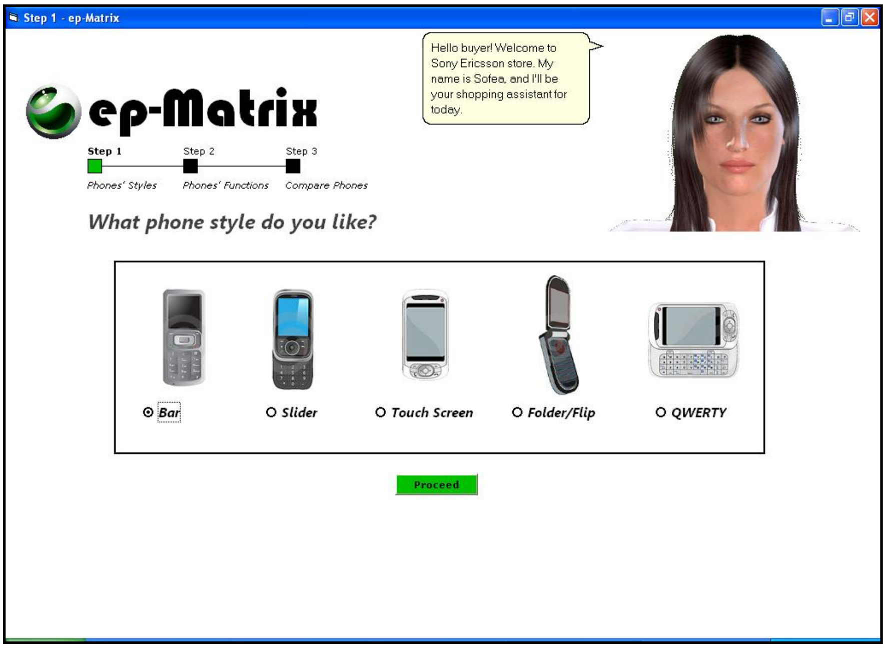
The prototype (ep-Matrix) is developed to demonstrate an example of making a purchasing decision of a mobile phone. A well know brand of mobile phone is used for three reasons; the convenience of getting all the required data, the familiarity factor among consumers and for the purpose of evaluation later on. Table 6.2 summarizes the features of ep-Matrix that is developed for this case study:

Table 6.2: Features of e^{p}-Matrix

ep-Matrix	
Alternatives filter	Mobile phone styles: Bar Slider* Touch Screen Folder/Flip QWERTY
Criteria	13 objective 9 subjective
Alternatives	Slider (6 models)
Feedback	Pop-up window, on-screen text, Interface agent

* *enabled in this prototype*

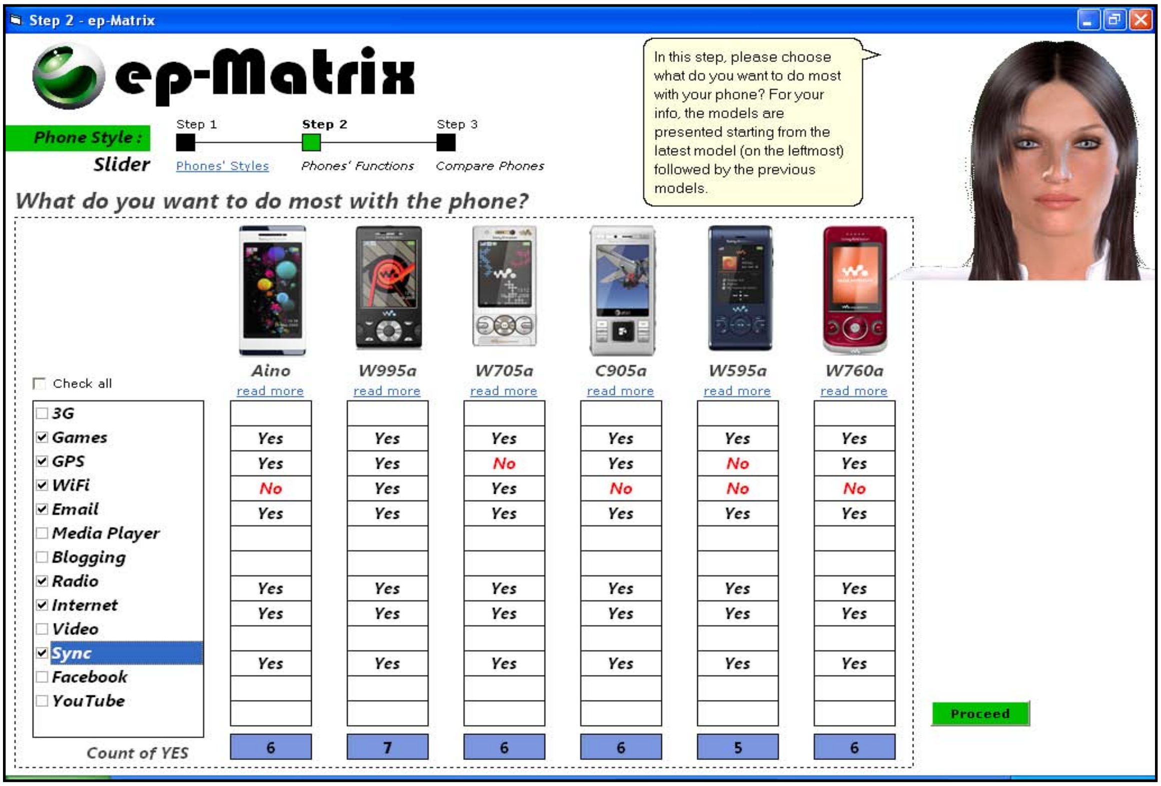
The first step of e^{p}-Matrix enables user to filter the available phone models based on preferred style (Figure 6.5). As it proceeds with the second step (Figure 6.6), users will make their selection of objective criteria to further filter the options (phone models) following the non-compensatory strategy (lexicographic process). The three highest scored models (see Figure 6.7) which pass most of the selected criteria will be ranked accordingly and the one in the highest rank will be set as the first reference (datum). Next, the three identified models from previous step will be compared to each other following the compensatory strategy (modified Pugh‘s method) based on preferred subjective criteria (Figure 6.8). The steps can be iterated in maximum 3 cycles where in each round the reference will be changed until each model will be a reference once. The dominance model from the 3 rounds will be suggested as the best selection. The following are screenshots of e^{p}-Matrix:



Hello buyer! Welcome to Sony Ericsson store. My name is Sofea, and I'll be your shopping assistant for today.

What phone style do you like?

Figure 6.5: Alternatives filtered categorically



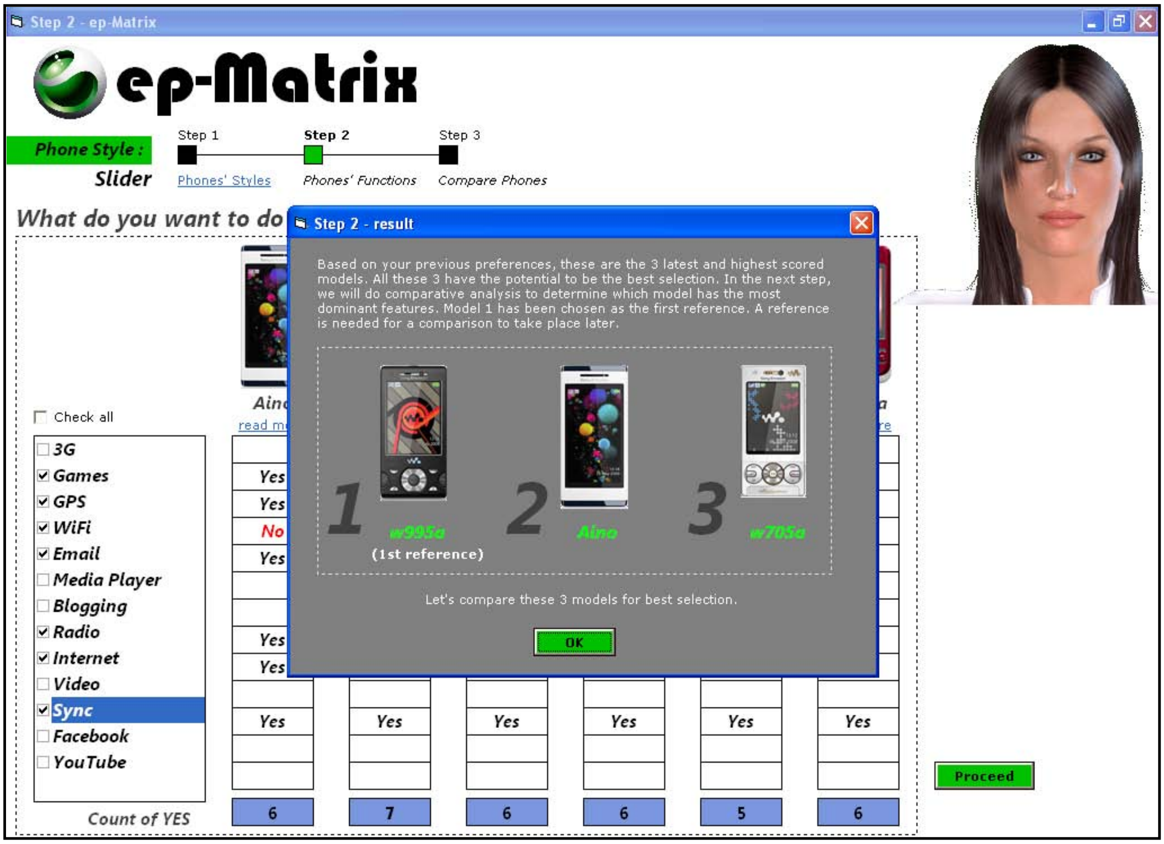
In this step, please choose what do you want to do most with your phone? For your info, the models are presented starting from the latest model (on the leftmost) followed by the previous models.

What do you want to do most with the phone?

□ Check all

	Aino read more	W995a read more	W705a read more	C905a read more	W595a read more	W760a read more
□ 3G						
☑ Games	Yes	Yes	Yes	Yes	Yes	Yes
☑ GPS	Yes	Yes	No	Yes	No	Yes
☑ WiFi	No	Yes	Yes	No	No	No
☑ Email	Yes	Yes	Yes	Yes	Yes	Yes
□ Media Player						
□ Blogging						
☑ Radio	Yes	Yes	Yes	Yes	Yes	Yes
☑ Internet	Yes	Yes	Yes	Yes	Yes	Yes
□ Video						
☑ Sync	Yes	Yes	Yes	Yes	Yes	Yes
□ Facebook						
□ YouTube						
Count of YES	6	7	6	6	5	6

Figure 6.6: The 13 objective criteria used in non-compensatory (lexicographic) process



What do you want to do

Step 2 - result

Based on your previous preferences, these are the 3 latest and highest scored models. All these 3 have the potential to be the best selection. In the next step, we will do comparative analysis to determine which model has the most dominant features. Model 1 has been chosen as the first reference. A reference is needed for a comparison to take place later.

1 w995a (1st reference) 2 Aino 3 w705a

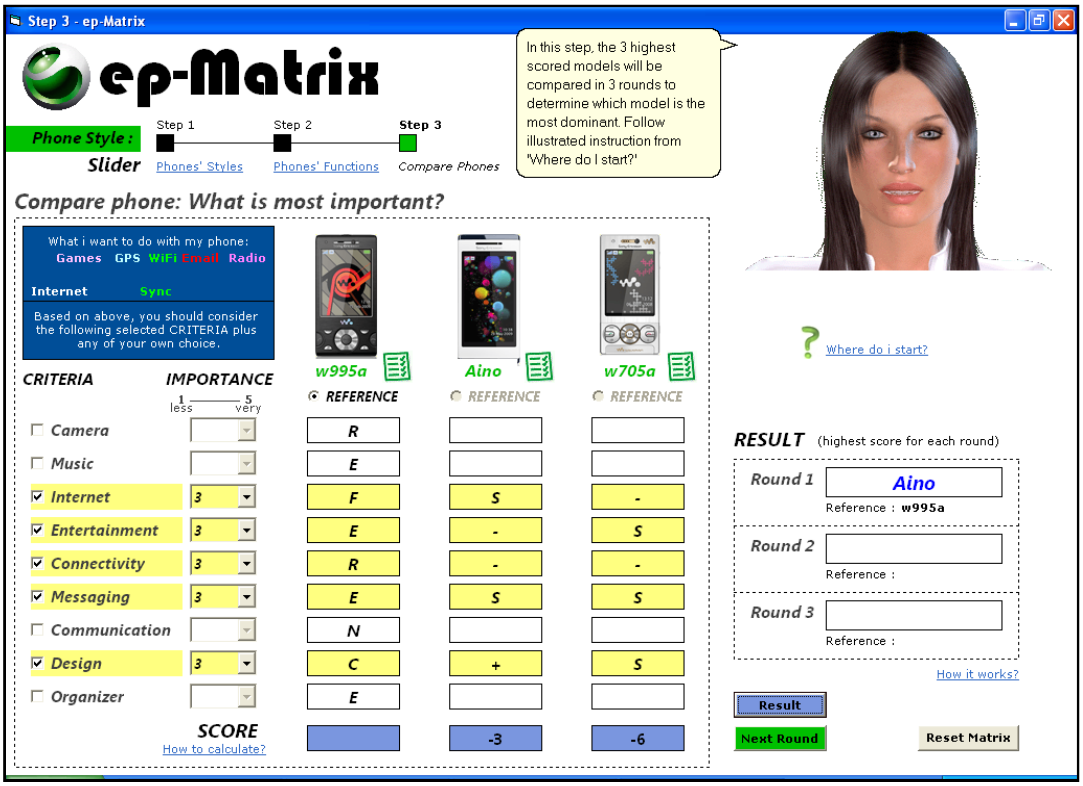
Let's compare these 3 models for best selection.

OK

□ Check all

	Ainc read m					a re
□ 3G						
☑ Games	Yes					
☑ GPS	Yes					
☑ WiFi	No					
☑ Email	Yes					
□ Media Player						
□ Blogging						
☑ Radio	Yes					
☑ Internet	Yes					
□ Video						
☑ Sync	Yes	Yes	Yes	Yes	Yes	Yes
□ Facebook						
□ YouTube						
Count of YES	6	7	6	6	5	6

Figure 6.7: Result obtained in non-compensatory process



Step 3 - ep-Matrix

ep-Matrix

Phone Style : Slider

Step 1 Phones' Styles — Step 2 Phones' Functions — Step 3 Compare Phones

In this step, the 3 highest scored models will be compared in 3 rounds to determine which model is the most dominant. Follow illustrated instruction from 'Where do I start?'

Compare phone: What is most important?

What i want to do with my phone:
Games GPS WiFi Email Radio
Internet Sync

Based on above, you should consider the following selected CRITERIA plus any of your own choice.

CRITERIA	IMPORTANCE (1 less — 5 very)	w995a ● REFERENCE	Aino ○ REFERENCE	w705a ○ REFERENCE
□ Camera		R		
□ Music		E		
☑ Internet	3	F	S	-
☑ Entertainment	3	E	-	S
☑ Connectivity	3	R	-	-
☑ Messaging	3	E	S	S
□ Communication		N		
☑ Design	3	C	+	S
□ Organizer		E		
SCORE How to calculate?			-3	-6

Where do i start?

RESULT (highest score for each round)

Round 1	Aino Reference : w995a
Round 2	 Reference :
Round 3	 Reference :

How it works?

Result

Next Round

Reset Matrix

Figure 6.8: The 9 subjective criteria used in compensatory process (modified Pugh's method)

VII. Helpfulness of PDADM Driven Prototypes

This study intends to investigate users' perception towards helpfulness of the PDADM driven prototypes in both case studies. In measuring helpfulness, quantitative data need to be gathered through an instrument. In addition to that, subjective input through interviews and observations might help enriching the collected data. To develop the instrument for measuring helpfulness, an elicitation work as summarized in Figure 7.1 was performed (Ariffin, 2009).

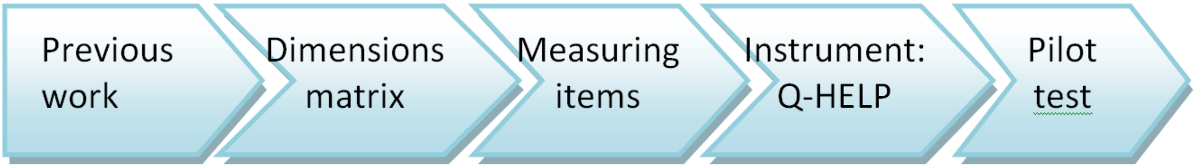


Figure 7.1: Summary of elicitation work

Figure 7.1 illustrates the processes involved in the instrument development; beginning with elicitation works to determine measuring items until the instrument is ready for pilot testing. The instrument was constructed based on the dimensions identified from elicitation work. Later, measuring items were added based on the reviewed literatures. Some modifications are made to the measuring items, in terms of rewording some items and repositioning some items into another dimension of the instrument. In measuring the helpfulness of the PDADM driven prototypes, this study is looking at four important dimensions; reliability, decision making effort, confidence, and decision process awareness. The instrument was then named as Q-HELP, which contains four dimensions: reliability, decision making effort, confidence, and decision process awareness

Table 7.1 illustrates the reliability of Q-HELP by each dimension. In the evaluation, respondents are required to rate the helpfulness level based on each dimensions using the seven point Likert scales; which are 1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = undecided, 5 = somewhat agree, 6 = agree and 7 = strongly agree. Respective measuring items can be seen in Table 7.2.

Table 7.1: Reliability of dimensions in Q-HELP

Dimensions	Cronbach Alpha value
Reliability	0.755
Decision making effort	0.689
Confidence	0.906
Decision process awareness	0.771

One hundred and seven respondents participated in the lab experiment; 63 of them were evaluated for the first case study where as 44 for the second case study. The experiment proceeded in two steps for each case study. In the first step, participants were required to accomplish the selection task aided by other tool or material. The main concern is to study the process that they went through before they can actually make a selection. In the second step, participants solved the same decision problem by making selection with the assistance of proposed PDADM driven prototypes in each case study.

Upon completion of both steps, participants were requested to answer 26 questions from all four dimensions of helpfulness in Q-HELP. The instrument recorded their perceptions and experiences of making a selection for the same decision problem in the experiment. Table 7.2 also depicts the mean responses for each item in Q-HELP answered by participants in respective case studies.

Table 7.2: Q-HELP items and mean responses by each item for each case study

Reliability	m^d-Matrix *n=63*	e^p-Matrix *n=44*
{*name of prototype*}* can be relied to function properly.	5.22	5.84
{*name of prototype*}* is suitable to my style of decision making.	5.02	5.43
{*name of prototype*}* is capable of helping me in making a choice.	5.25	5.80
{*name of prototype*}* provides the help that I need to make a selection.	5.33	5.75
{*name of prototype*}* provides the advice that I require to make my decision.	5.08	5.64
I would use {*name of prototype*}* if I were attempting to make a choice that is ―good enough” but not necessarily the best.	4.95	5.82
{*name of prototype*}* is suitable even during limited time to make a decision.	5.03	5.82
Group Mean A	5.13	5.73
Decision making effort		
It was very time consuming to choose a {item} from the available options.	4.81	5.39
It was very difficult to choose a {item} from the available options.	4.43	5.27
{*name of prototype*}* allowed me to carefully consider the decision made.	5.35	5.84
The decision process in {*name of prototype*}* is logical to me.	5.30	6.14
The decision process in {*name of prototype*}* is simple to me.	5.19	5.91
I understand how decision process in {*name of prototype*}* works.	5.17	5.70
I found it very easy to interpret the decision justification provided by {*name of prototype*}*.	5.06	5.77
Group Mean B	5.04	5.72
Confidence		
I am satisfied with the recommended solution.	5.27	5.75
The recommended solution reflects my initial preferences.	5.16	5.61
I am confident that I am able to make selection with {*name of prototype*}*.	5.17	5.86
I am confident that I can justify the selection that I made with {*name of prototype*}*.	5.17	5.93
I feel that the problem in making selection is solved.	5.05	5.45
I am very pleased with my experience using {*name of prototype*}*.	5.48	5.77
Group Mean C	5.22	5.73
Decision process awareness		
{*name of prototype*}* makes me realize I cannot get everything from just one alternative.	5.44	5.93
{*name of prototype*}* is an aid for me in clarifying what I want.	5.27	5.84
{*name of prototype*}* shows my subconscious decision process.	5.11	5.73
{*name of prototype*}* helps me not to be easily influenced by others in making selection.	5.29	5.98
{*name of prototype*}* makes me more independent of others in making a selection.	5.22	6.00
I learned a lot about the problem using {*name of prototype*}*.	5.48	6.00
Group Mean D	5.30	5.91

**replaced with m^d-Matrix or e^p-Matrix based on respective case studies*

VIII. Results

As mentioned earlier, the instrument used in evaluating the helpfulness of the PDADM driven prototypes is looking at four important dimensions; reliability, decision making effort, confidence, and decision process awareness. Table 7.2 presents means of responses to the items in measuring the helpfulness of the prototypes in both case studies.

Questions A1 to A7 are used to assess the user‗s perceptions on reliability of the prototypes. For case study 1, the group mean score of items in dimension A was 5.13, indicating moderately high perception on reliability. In case study 2, the group mean score of the same items was 5.73, indicating high level of reliability.

Question B1 to B7 are used to assess the user‗s perceptions on effort invested in the decision making process with the assistance of PDADM driven prototypes. For case study 1, the group mean score for items in dimension B was 5.04, signifying moderately high perception on decision making effort among respondents. As for case study 2, the group mean score of the same items was 5.72, indicating high perception on the decision making effort.

Question C1 to C6 are used to assess the confidence level of respondents in solution and procedure applied in the decision aids. In case study 1, the group mean score was 5.22, representing moderate confidence level among respondents. As for the second case study, the group mean score was 5.73, indicating higher confidence level among respondents after using the PDADM driven prototypes.

For the last dimension of the instrument, six items (items D1 to D6) have been asked to the respondents in order to measure their perception on decision process awareness. In case study 1, the group mean score of the last six items in Q-HELP was 5.30, representing moderate perception score on decision process awareness among respondents. For case study 2, the group mean score was 5.91, signifying high perception score on decision process awareness.

From the analysis above and as can be summarized in Figure 6.9 , generally the mean scores of each dimension fall under category moderately high or high, indicating that participants were incline to perceive the use of PDADM driven prototypes as helpful even in different personal decision situations. In both prototypes, participants rated highly on decision process awareness, this is followed by their perceived confidence and reliable in the decision aids.

Upon further analysis, participants responded highly on the items under reliability and confidence as depicted in Figure 6.10 and 6.11. Therefore, it can be concluded that both decision aids:

i. provide the help that participants needed to make a selection,
ii. can be relied to function properly
iii. are capable of helping participants in making a choice

Also, the participants were:

i. very pleased with their experience using the decision aids
ii. confident that they can justify the selection that have been made with the decision aids
iii. satisfied with the recommended solution
iv.

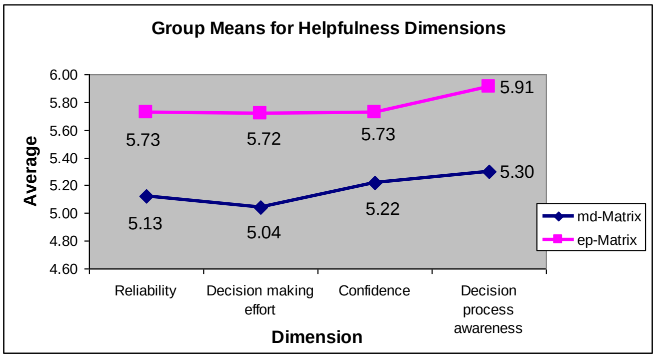


Figure 6.9: Group means for helpfulness dimensions

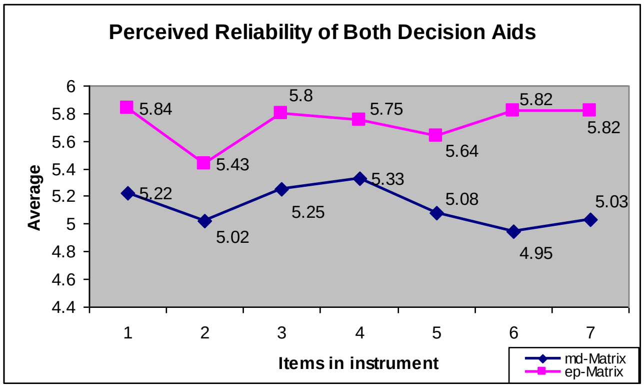


Figure 6.10: Perceived reliability of m^{d}-Matrix and e^{p}-Matrix

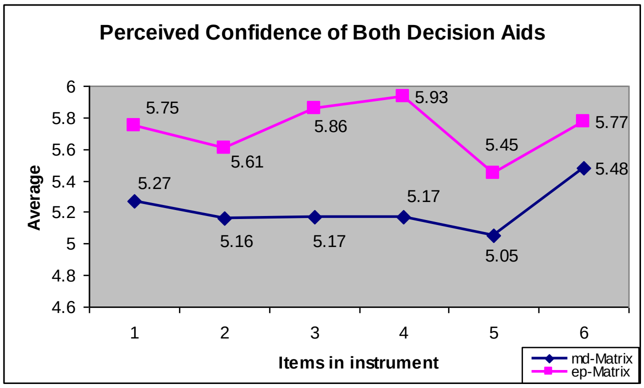


Figure 6.11: Perceived confidence in m^{d}-Matrix and e^{p}-Matrix

IX. Conclusion

Despite the existence of various computerized decision aids, decision maker perceptions of the ideal decision strategy and technique have not been subjected to systematic investigation. In doing so, this study seeks to contribute the following, along achieving the previously stated objectives:

i. In general, this study will contribute to decision making area as well as cross-disciplinary area which is related to the decision situation
ii. A proposed decision making model for personal decisions with emphasis on the non-expert use.

iii. Two prototypes which utilizing the proposed decision model in two different situations; purchasing decision and educational decision.

iv. Algorithms of the developed prototypes.

v. Instruments to measure users' perceived helpfulness of the prototypes.

vi. A comparative analysis of five decision strategies which provides research basis for related future studies.

X. REFERENCES

1) Adam, F. and Humphreys, P. (2008). Encyclopedia of Decision Making and Decision Support Technologies. Idea Group Inc.

2) Alidrisi, M.M. (1987). Use of multi attribute utility theory for personal decision making. *International Journal of Systems Science*, 18(12), 2229—2237.

3) Al-Shemmeri, T., Al-Kloub, B. and Pearman, A. (1997). Model Choice in Multicriteria Decision Aid. *European Journal of Operational Research*, 97, 550-560.

4) Afonso, A.P., Regateiro, F.S., and Silva, M.J. (1998). *Dynamic Channels: A New Development Methodology for Mobile Computing Applications.* Retrieved, Jan 22, 2007, from http://www.di.fc.ul.pt/biblioteca/tech-reports.

5) Ariffin, A.M. (2009). *Conceptual Design Model of Reality Learning Media (RLM): Towards Entertaining and Fun Electronic Learning Materials (eLM)* (Ph.D. Dissertation, Universiti Utara Malaysia)

6) Arsham, H. (2004). *Decision Making: Overcoming Serious Indecisiveness*. Retrieved March 10, 2009 from http://home.ubalt.edu/ntsbarsh/opre640/partXIII.htm.

7) Atkinson, C. and Olla, P. (2004). Developing a wireless reference model for interpreting complexity in wireless projects. *Industrial Management & Data Systems*, 104, 262-272.

8) Avison, D.E. and Fitzgerald, G. (1990). *Information Systems Development: Methodologies, Techniques and Tools*. London: Blackwell.

9) Bahl, H.C. and Hunt, R.G. (1984). Decision-Making Theory and DSS Design. *Data Base*, 15(4), 10-14.

10) Baker, D., Bridges, D., Hunter, R., Johnson, G., Krupa, J., Murphy, J. and Sorenson, K. (2002) Guidebook to Decision-Making Methods, WSRC-IM-2002-00002, Retrieved from Department of Energy, USA website: http://emi-web.inel.gov/Nissmg/Guidebook_2002.pdf.

11) Bell, D.E., Raiffa, H., and Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision making. In D. Bell, Raiffa, H., and A. Tversky (Eds.), *Decision making: descriptive, normative, and prescriptive interactions* (pp. 9-32). Cambridge: Cambridge University Press.

12) Bertini, E., Gabrielli, S., and Kimani, S. (2006). Appropriating and Assessing Heuristics for Mobile Computing. *Proceedings of the working Conference on Advanced Visual Interfaces AVI'06*, Venezia, Italy. 119-126.

13) Bridgman, P. W. (1922). *Dimensional analysis*. New Haven, CT: Yale University Press.

14) Brown, R. (2008). Decision Aiding Research Needs. In. Adam F. and Humphreys P. (Eds.), *Encyclopedia of Decision Making and Decision Support Technologies* (pp. 141-147). IGI Global.

15) Bronner, F. & de Hoog, R. (1982). Non-Expert Use of a Computerized Decision Aid. In Humphreys, P., Svenson, O. and Anna Vári, A. (Eds.), *Analysing and Aiding Decision Processes* (pp. 281-299). North Holland: Amsterdam.

16) Christ, P. (2008). *KnowThis: Marketing Basics*. KnowThis Media.

17) Collins, T.R., Rossetti, M.D., Nachtmann, H.L. & Oldham J.R. (2006). The use of multi-attribute utility theory to determine the overall best-in-class performer in a benchmarking study. *Benchmarking: An International Journal*, 13, 431-446.

18) Cosier, R.A. and Dalton, D.R. (1986). The Appropriate Choice and Implementation of Decision Strategies. *Journal of Industrial Management & Data Systems*, 86(3/4), 18-21. Abstract retrieved from http://www.emeraldinsight.com/10.1108/eb057436

19) Dyer, J.S., Fishburn, P.C., Steuer, R.E., Wallenius, J. and Zionts, S. (1992). Multiple Criteria Decision Making, Multiattribute Utility Theory: The Next Ten Years. *Management Science*, 38(5), 645-654.

20) Easwaran, Kenny (2009). *Dominance-based Decision Theory.* Unpublished manuscript. Retrieved from http://www.ocf.berkeley.edu/~easwaran/papers/decision.pdf

21) Einhorn, H.J. and Hogarth, R.M. (1981). Behavioral Decision Theory: Process of Judgment and Choice. *Annual Reviews Psychology*, 32, 53-88.

22) Fishburn, P.C. (1967). Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments. *American Society of Operations Research (ORSA)*, Baltimore, MD: U.S.A.

23) Gillespie, M. (2007). *Resource Guide for the UMPC Software Developer*. Intel.com

24) Häubl, G. and Trifts, V. (2000). Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science*, 19(1), 4-21.

25) Hayes, C.C. & Akhavi, F. (2008). Creating Effective Decision Aids for Complex Tasks. *Journal of Usability Studies*. 3 (4), 152 - 172.

26) Heikkinen, M.T. and Still, J. (2005). Business Networks and New Mobile Service Development. *Proceedings of the International Conference on Mobile Business (ICMB'05)*. 144 -151.
27) Heyes, I.S. (2002). *Just Enough Wireless Computing*. Upper Saddle River, NI: Prentice Hall.
28) Işıklar, G. and Büyüközkan, G. (2007). Using a multi-criteria decision making approach to evaluate mobile phone alternatives. *Computer Standards & Interfaces*, 29, 265-274.
29) Jungermann, H. (1980). Speculations about Decision Theoretic Aids for Personal Decision Making. In Acta Psychologica 45 (pp. 7-34). North Holland.
30) Keeney, R. and Raiffa, H. (1993). *Decisions with Multiple Objectives : Preference and Value Tradeoffs*, Cambridge University Press, Cambridge.
31) König-Ries, B. (2009). Challenges in Mobile Application Development. it – *Information Technology*, 51(2), 69-71.
32) Law, W. S. (1996). *Evaluating imprecision in engineering design* (Ph.D. Dissertation, California Institute of Technology, Pasadena, California).
33) Linkov, I., Varghese, A., Jamil, S., Seager, T.P., Kiker, G. and Bridges, T. (2004) Multi-criteria decision analysis: A framework for structuring remedial decisions at the contaminated sites, In: Linkov, I. and Ramadan, A.B. (Eds.), *Comparative Risk Assessment and Environmental Decision Making* (pp. 15-54). New York: Springer.
34) March, S.T. and Smith, G. (1995). Design and Natural Science Research on Information Technology. *Decision Support Systems*, 15(4), 251-266.
35) McGuire, R. (2002). Decision Making. *The Pharmaceutical Journal*. 269, 647-649.
36) Miller, D.W., & Starr, M.K. (1969). *Executive decisions and operations research*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
37) Naude, P., Lockett, G. and Holms, K. (1997). A Case Study of Strategic Engineering Decision Making Using Judgmental Modeling and Psychological Profiling. *Transactions on Engineering Management*, 44(3), 237-247.
38) Offermann, P., Levina, O., Schonherr, M. and Bub, U. (2009). Outline of a Design Science Research Process. *Proceedings of DESRIST'09*, Malvern, PA: USA.
39) Payne, J.,Bettman, J. and Johnson, E. (1993). *The Adaptive Decision Maker*. Cambridge University Press.
40) Power, D.J. (1998). *Designing and Developing a Computerized Decision Aid - A Case Study*. Retrieved December 10, 2009 from http://dssresources.com/papers/decisionaids.html.
41) Pugh, S. (1990). *Total Design: Integrated Methods for Successful Product Engineering*. Great Britain: Addison Wesley.
42) Rich, P. (1999). *A Process for Effective Decision Making*. Retrieved 5 April 2009 from http://www.selfhelpmagazine.com/article/decision-making
43) Saaty, T.L. (1977). A Scaling Method for Priorities in Hierarchical Structures. *Journal of Mathematical Psychology*, 15, 57-68.
44) Saaty, T.L. (1994). *Fundamentals of Decision Making and Priority Theory with the AHP*. Pittsburgh, PA: RWS Publications.
45) Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63, 129–138.
46) Soelberg, P.O. (1967). Unprogrammed Decision Making. *Industrial Management Review*, 8, 19-29.
47) Todd, P. & Benbasat, I. (1991). An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies. *Information Systems Research*, 2(2), 87-115.
48) Triantaphyllou, E. (2000). Multi-Criteria Decision Making Methods: A Comparative Study. Norwell, MA: Springer.
49) Ullman, D.G. (2002). *The Ideal Engineering Decision Support System*. Retrieved March 10, 2009 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.1827&rep=rep1&type=pdf
50) Westaby, J.D. (2005). Behavioral reasoning theory: Identifying new linkages underlying intentions and behavior. *Organizational Behavior and Human Decision Processes*, 98, 97-120.
51) Wooler, S. (1982). A Decision Aid for Structuring and Evaluating Career Choice Options. *Journal of Operational Research Society*, 33(4), 343-351.
52) Zanakis, S.H., Solomon, A., Wishart, N. and Dublish, S. (1998). Multi-attribute decision making: A simulation comparison of select methods. *European Journal of Operational Research,* 107, 507-529.
53) Zannier, C, Chaisson, M. and Maurer, F. (2007). A model of design decision making based on empirical results on interviews with software designers. *Information and Software Technology*, 49, 637-653.
54) Zsambok, C.E., Beach, L.R. & Klein, G. (1992) *A Literature Review of Analytical and Naturalistic Decision Making*. Final technical report, Fairborn, OH: Klein Associates Inc.

Extraction of Interesting Association Rules using GA Optimization

GJCST Computing Classification
G.1.6, H.2.8, H.1.1

Virendra Kumar Shrivastava[1] Dr. Parveen Kumar[2] Dr. K. R. Pardasani[3]

***Abstract*-Association rule mining is a process of discovering interesting and unexpected rules form very large databases. Discovery of association rules at primitive-level is called single-level association rules or primitive-level association rules. However, mining association rules at multi-level may lead to the discovery of more specific and useful knowledge from dataset. Mining of Multi-level Association Rules (MLAR) are not useful until it can be used to improve decision making process. The main hurdle in this process is the number of rules grows exponentially with the number of items. Support and confidence limit the level of interestingness of the generated rules. However, the challenge arises in selection of interesting rules from the set of rules. In this research paper, we endeavor to optimize the rules generated by FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases using genetic algorithm. The rules may optimize using measures like support, confidence factor, interestingness and completeness.**

Keywords-Discovery of multi-level association rules, interestingness, completeness and Genetic Algorithm.

I. Introduction

Techniques of association rule mining can be used to discover unknown or hidden correlation between items found in the database of transactions. An association rule [1,2,4,5,8,11,12] is a rule, which implies certain association relationships among a set of attributes (such as 'occurs together' or 'one implies to other') in a database. Apriori [5] is the most popular and influent algorithm to find all the frequent itemsets. It is proposed by Agrawal and Srikant in 1994. It is also called the level-wise algorithm. Multi-level association rules mining involves items at different level of abstraction. For many applications, it is difficult to find strong association among data items at low or primitive level of abstraction. Associations discovered at higher levels may represent common sense knowledge. Multi-level techniques find rules that are hidden or impossible to mine when searching at the primitive-level. This is because the conventional algorithms neglect several items from analysis that do not appear often enough to be considered significant. For example, joystick may not be purchased frequently, and therefore omitted from association rules. However, by using concept hierarchies, we can place joystick in a larger category containing mouse, pen drive, mouse pad, and joystick etc., called computer accessories. Thus, indirectly include their name in association mining process. To discover multilevel association rules, one need to provide (i) data at multi-levels of abstraction and (ii) efficient methods for multi-level rule mining. Researchers have given some methods for MLARM [5,7,10,15]. In this research, we considered FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases [3]. But all rules generated by this method may not be interesting. The main hurdle in this process is the number of rules grows exponentially with the number of items. The rules may optimize using measures like support, confidence factor, interestingness and completeness. The prime objective of this paper is to find interesting rules (of high predictive accuracy) from given data set using optimization of Genetic Algorithm.

This paper is organized as follows. The section two describes the basic concepts related to the multiple level association rules. In section three, discuss the Genetic Algorithm. Section four describes the results, and finally we conclude our research work in section five.

II. Multiple-level Association Rules

The multi-level association rule mining utilizes a concept hierarchy. This hierarchy represents the relationship among different concept levels. For example, in figure 1, we represent a concept hierarchy that one might find in a typical electronics / computer sales house. To explore the mining of association rules from a largeset of transaction data, let assume that the database contains:

i. an item data set which contains the description of each item in I in the form of $<A_i, description_i>$, where A_i Є I, and

ii. a transaction data set T , which consists of a set of transactions $<T_i, \{A_p, . . .,A_q\}>$, where T_i is a transaction identifier and A_i Є I (for i . p, . . . , q).

Definition: A pattern or an itemset A, is one item Ai or a set of conjunctive items Ai ^…^ Aj, where Ai, . . . , Aj Є I. The support of a pattern A in a set S, s(A/S), is the number of transactions (in S) which contain A versus the total number of transactions in S. The confidence of A => B in S, c(A => B/S), is the ratio of s(A ^ B/S) versus s(A/S), i.e., the probability that pattern B occurs in S when pattern A occurs in S.To generate relatively frequent occurring patterns and reasonably strong rule implications, one may specify two thresholds: minimum support s´, and minimum confidence c´. Observe that, for finding multiple-level association rules,

About[1]-Department of Computer Engineering, Singhania University, Pacheri Bari, (Rajsthan), India (e-mail- vk_shrivastava@yahoo.com)
About-[2]Department of Computer Science & Engineering,MIET, Meerut (U. P.), India (e-mail-pk223475@yahoo.com)
About-[3]Dept. of Maths & MCA Maulana Azad National Inst. Of Tech., Bhopal, (M. P.) India (e-mail-kamalrajp@hotmail.com)

different minimum support and/or minimum confidence can be specified at different levels.

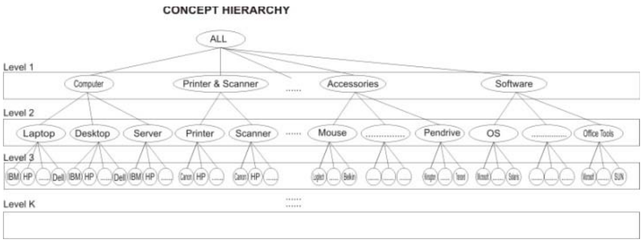


Figure 1: concept hierarchy

We attempt to optimize the rules generated by FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases [3]. This rule generation method uses hierarchy information encoded in transaction table instead of the original transaction table. This is because; first a data mining query is usually in relevance to only a portion of the transaction database, such as computer, printer etc instead of all the items. Thus, it is useful to first collect the

relevant set of data and then work repeatedly on th`e task related set. Second, encoding can beperformed during the collection of task related data and thusthere is no extra encoding pass required. Third, an encoded string, which represents a position in a hierarchy, requires lesser bits than the corresponding bar code. Thus,it is often beneficial to use an encoded table. For example, theitem `IBM Desktop Computer' is encoded as `111' in whichthe first character, `1', represents `Computer' at level-1, the second, `1',for `laptop (computer)' at level-2, and the third, `1', for the brand`IBM' at level-3. Repeated items (i.e., itemswith the same encoding) at any level will be treated as oneitem in one transaction.

The FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases consists of two main phases. Phase one is the construction of a modified Frequent Pattern tree. Phase two is the repetitive building of small data structures, the actual mining for these data structures, and their release. The association rules are generated at multiple- level using the frequent patters at related concept level.

III. Genetic Algorithm

Genetic algorithm (GA) was first developed by John Holland at university of Michigan in 1975. It incorporates Darwinianevolutionary theory with sexual reproduction. GA is stochastic search algorithm modeledon the process of natural selection, which underlines biological evolution. GA has beensuccessfully applied in many search, optimization, and machine learning problems. A group of individuals called population, is stored and modified during each iteration of the algorithm. In GA's iterations are referred to as generations. GA processes generations by generating new populations of strings from old ones.Every string is the encoded binary, real etc., version of a candidate solution. An evaluationfunction associates a fitness measure to every string indicating its fitness for the problem.Standard GA[9,13] apply genetic operators such selection, crossover and mutation on an initiallyrandom population in order to compute a whole generation of new strings. It generates solution for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Therefore, the quality of solutions in successive generations improves. The GA process is terminated when an acceptable or optimum solutions is found.

The function of GA is as follows:

•*Selection*deals with the probabilistic survival of the fittest, in that more fit chromosomes arechosen to survive. Where fitness is a comparable measure of how well a chromosome solves theproblem at hand.

• *Crossover*specifies how the genetic algorithm combines two individuals, or parents, to form a crossover child for the next generation.

• *Mutation*alters the new solutions so as to add stochasticity in the search for better solutions. This is the chance that a bit within a chromosome will be flipped (o becomes 1 and vice versa).

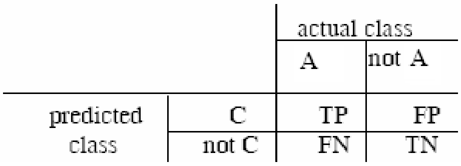
*Fitness function:*Ideally the discovered rules shouldhave (a) a high predictive accuracy; (b) be comprehensible; and (c) be interesting.The accomplishment of a genetic algorithm is directly linked to the accuracy of the fitness function. The fitness function should be customized to the specific search spaces. We take a fitness function that considers major issues in evaluating an individual against its search space. The fitness of a population is the sum of the individual fitness values of that population. The fitness function is the primary performance sink in a genetic algorithm, because this is the place that the underlying data must be accessed. Therefore, optimization should be considered wherever possible. The general structure of a rule is defined as:

IV. If antecedent than consequent

Let a rule be of the form:

IF A THEN C,

Where A is the antecedent (a conjunction of conditions) and C is the consequent (predicted class). The predictive performance of a rule can be summarized by a 2 x 2 matrix, sometimes called a confusion matrix, as depicted in the following figure 2:



		actual class	
		A	not A
predicted class	C	TP	FP
	not C	FN	TN

Figure 2: Confusion Matrix for a rule

The abbreviation and meaning of the labels used in the confusion matrix have the following meaning:

TP = True Positives = Number of examples satisfying A item set and item set C

FP = False Positives = Number of examples not satisfying item set A but satisfying item set C

FN = False Negatives = Number of examples satisfying item set A but not satisfying item set C
TN = True Negatives = Number of examples not satisfying A nor C
It is clear that the higher the values of TP and TN, and the lower the values of FP and FN, the better the rule.
Interestingness Factor (INF) = TP/(TP+FP)
Now measure the predictive accuracy of a rule by taking into account not only its INF but also a measure of how "complete" the rule is, i.e. what is the proportion ofexamples having the predicted class C that is actually covered by the rule antecedent. The rule completeness factor measure, denoted CF, is computed as:
Completeness Factor (CF) = TP / (TP+FN)
In order to combine the INF and CF measures one can define a fitness function such as:
Fitness = INF x CF
Although this fitness function does a good job in evaluating predictive performance, it has nothing to say about the comprehensibility of the rule. This fitness function can be extended (or any other focusing only on the predictive accuracy of the rule) with a rule comprehensibility measure in several ways. A simple approach is to define a fitness function such as
Fitness = w1 × (INF × CF) + w2 × S
Where, S is a measure of rule simplicity. The S values lie between [0, 1] and w1 and w2 are user-defined weights. In general, its value is inversely proportional to the number of conditions in the rule antecedent – i.e., the shorter the rule, the simpler it is.

V. Optimization Methodology

We applied GA over the rules generated from FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases [3]. Optimization [14,16] does not mean maximization or minimization. Optimization is means to get the most feasible solution or utilization of the available methodology for their best uses. The following genetic algorithm is used to optimize (i.e. finding interesting relationships) rules at level l.

1. Create random population of n chromosomes.
2. Calculate fitness for each chromosome in the population
3. Selection – based on fitness function
4. Apply Cross-over and mutation on the selected members
5. Accept or reject new one
6. Replace old with new population
7. Test problem criterion
8. Repeat step 2-7 until criterion is satisfied

The genetic algorithm mechanism can be explained with the following flow chart (as given in figure 3).

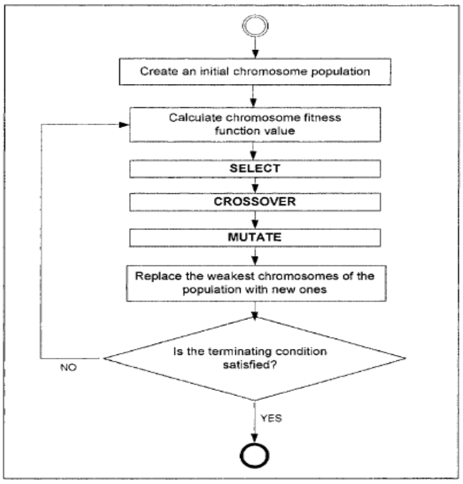

Create an initial chromosome population
Calculate chromosome fitness function value
SELECT
CROSSOVER
MUTATE
Replace the weakest chromosomes of the population with new ones
Is the terminating condition satisfied?
NO
YES

Figure 3: The genetic algorithm mechanism

VI. Results

We have applied GA to the rules obtained by the FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules. We have implemented genetic algorithm for optimization in MATLAB [15]. We have used computer user's feedback dataset to test the effectiveness of proposed algorithm. We have used MATLAB 7.0 and tested all experiments on a Dell Laptop with Intel® Core™2 Duo 2.0 GHz processor and 2.00 GB of main memory using Microsoft Windows XP operating system. The following table shows the values of TP, FP,
FN, interestingness, completeness and fitness of the rule.

TP	FP	FN	INF	CF	Fitness
20	12	7	0.625	0.741	.0463
34	23	16	0.596	0.68	0.405
35	30	14	0.538	0.714	0.384
8	32	14	0.2	0.364	0.072
25	34	12	0.424	0.676	0.286
16	29	16	0.356	0.5	0.177

Table 1: obtained values of TP, FP, FN, interestingness, completeness and fitness function.

Given below, Table 2 describes parameters used in the genetic algorithm implementation

Selection	Tournament, size = 3
Crossoverprobability	0.1
MutationProbability	= 0.005
Fitness function	Discussed in section 3.
GA population	100

Table 2: Genetic algorithm parameters

Our experimental results show that the optimized rules have a high interestingness and completeness.

VII. Conclusion

In this research work, we have used multi-level association rules generated by FP-tree and COFI Based Approach. All generated rules are not interesting. We have apllied Generic Algorithm to optimize the association rules. We obtain a fitness function for the task of optimization and find the optimum solutions that are interesting rules. It extracts interesting rules with predictive accuracy.

VIII. References

1) R. Agrawal, T. Imielinski, and A. Swami.. "Mining association rules between sets of items in large databases". In Proceedings of the 1993 ACM SIG-MOD International Conference on Management of Data, pages 207-216, Washington, DC, May 26-28 1993.
2) R Srikant, Qouc Vu and R Agrawal. "Mining Association Rules with Item Constrains". IBM Research Centre, San Jose, CA 95120, USA.
3) Virendra Kumar Shrivastava, Parveen Kumar, K. R. Pardasani, "FP-tree and COFI Based Approach for Mining of Multiple Level Association Rules in Large Databases". International Journal of Computer Science and Information Security (IJCSIS), pp 273 – 279, USA, 2010.
4) Ashok Savasere, E. Omiecinski and Shamkant Navathe "An Efficient Algorithm for Mining Association Rules in Large Databases". Proceedings of the 21st VLDB conference Zurich, Swizerland, 1995.
5) R. Agrawaland R, Shrikanth, "Fast Algorithm for Mining Association Rules". Proceedings Of VLDB conference, pp 487 – 449, Santigo, Chile, 1994.
6) Arun K Pujai "Data Mining techniques". University Press (India) Pvt. Ltd. 2001.
7) Jiawei Han and Yongjian Fu "Discovery of Multiple-Level Association Rules from Large Databases". Proceedings of the 21st VLDB Conference Zurich, Swizerland, 1995.
8) J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufman, San Francisco, CA, 2001.
9) Melanie Mitchell, 1996. An Introduction to genetic Algorithms, PHI.
10) Jiawei Han and Yongjian Fu "Discovery of Multiple-Level Association Rules from Large Databases". IEEE Trans. on Knowledge and Data Eng. Vol. 11 No. 5 pp 798-804, 1999.
11) J. Han, J, Pei and Y Yin. "Mining Frequent Patterns Without Candidate Generation". In ACM SIGMOD Conf. Management of Data, May 2000.
12) Y. Wang, Y. He and J. Han. "Mining Frequent Item Sets Using Support Constraints." In Proceedings 2000 Int Conference VLDB'00, Carid; Egypt, Sep. 2000, Page 43-52.
13) Anandhavalli M., Suraj Kumar Sudhanshu, Ayush Kumar and Ghose M.K. "Optimized association rule mining using genetic algorithm" In Advances in Information Mining, Volume 1, Issue 2, 2009, pp-01-04.
14) Manish Saggar, Ashish Kumar Agrawal, and Abhimanyu Lad "Optimization of Association Rule Mining using Improved Genetic Algorithms". In 2004 IEEE proceedings.
15) R. S. Thakur, R. C. Jain and K. R. Pardasani " Fast Algorith for mining multi-level association rules in large databases". Asian Journal of International Management 1(1):19-26, 2007.
16) Chipperfield, A.J. and P.J. Fleming, 1995. "The MATLAB genetic algorithm toolbox". Department of Automatic Control and Systems Engineering, University of Sheffield, PO Box 600, Mappm Street, Sheffield, England SI 4DU, From IEE Colloquium on Applied Control Techniques Using MATLAB, Digest.

Global Journals Guidelines Handbook 2010

www.GlobalJournals.org

FELLOW OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (FICCT)

- FICCT' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FICCT" can be added to name in the following manner e.g. **Dr. Andrew Knoll, Ph.D., FICCT, Er. Pettor Jone, M.E., FICCT**
- FICCT can submit two papers every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free unlimited Web-space will be allotted to 'FICCT 'along with subDomain to contribute and partake in our activities.
- A professional email address will be allotted free with unlimited email space.
- FICCT will be authorized to receive e-Journals - GJCST for the Lifetime.
- FICCT will be exempted from the registration fees of Seminar/Symposium/Conference/Workshop conducted internationally of GJCST (FREE of Charge).
- FICCT will be an Honorable Guest of any gathering hold.

ASSOCIATE OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (AICCT)

- AICCT title will be awarded to the person/institution after approval of Editor-in-Chef and Editorial Board. The title 'AICCT can be added to name in the following manner:
  eg. **Dr. Thomas Herry, Ph.D., AICCT**
- AICCT can submit one paper every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free 2GB Web-space will be allotted to 'FICCT' along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted with free 1GB email space.
- AICCT will be authorized to receive e-Journal GJCST for lifetime.
- A professional email address will be allotted with free 1GB email space.
- AICHSS will be authorized to receive e-Journal GJHSS for lifetime.

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJCST for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two-peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.

Online Submission: There are three ways to submit your paper:

(A) (I) Register yourself using top right corner of Home page then Login from same place twice. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal from "Research Journals" Menu.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer (Although Mozilla Firefox is preferred), then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org as an attachment.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION **(Must be strictly followed)**

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Times New Roman.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be two lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.

Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global Journals are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals and Editorial Board, will become the copyright of the Global Journals.

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments. Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads: Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also. Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals, ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: *Please make these as concise as possible.*

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals.

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals (Publication Prior to Print)

The Global Journals are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

Informal Tips for writing a Computer Science Research Paper to increase readability and citation

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

Techniques for writing a good quality Computer Science Research Paper:

1. Choosing the topic- In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table

- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.
- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)

- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.

You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described.
Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

Administration Rules Listed Before Submitting Your Research Paper to Global Journals

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals:

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.
- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation) by Global Journals

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals.

Topics	Grades		
	A-B	**C-D**	**E-F**
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

Index

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350