January - July, 2009 Vol. 1 No. 1 ISSN 0973-5658

BVICAM's IJIT

BVICAM's
International Journal of Information Technology

CONTENTS

BHARATI VIDYAPEETH PUNE

Bharati Vidyapeeth's
Institute of Computer Applications and Management
A-4, Paschim Vihar, Rohtak Road, New Delhi-63

Email : bijit@bvicam.ac.in, Website : http://www.bvicam.ac.in

Volume 1, Number 1 (Inaugural Issue)

January – July, 2009

BVICAM's International Journal of Information Technology (BIJIT) is a bi-annual publication of Bharati Vidyapeeth's Institute of Computer Applications and Management, A-4, Paschim Vihar, New Delhi – 110063.

ISSN 0973 – 5658

Disclaimer
The opinions expressed and figures provided in this Journal; BIJIT, are the sole responsibility of the authors. The publishers and the editors bear no responsibility in this regard. Any and all such liabilities are disclaimed

All disputes are subject to Delhi jurisdiction only.

Address for Correspondence:

The Director,
Bharati Vidyapeeth's
Institute of Computer Applications and Management,
A-4, Paschim Vihar, New Delhi – 110063 (INDIA).
Tel./Fax: 91 – 11 – 25275055 E-Mail: bijit@bvicam.ac.in
Visit us at www.bvicam.ac.in

Published and printed by Prof. M. N. Hoda, Director, Bharati Vidyapeeth's Institute of Computer Applications and Management, A-4, Paschim Vihar, New Delhi – 63 (INDIA).
Tel. / Fax: 91 – 11 – 25275055, E-Mail: bijit@bvicam.ac.in

BVICAM's
International Journal of Information Technology (BIJIT)

BIJIT is a bi-annual publication of
Bharati Vidyapeeth’s
Institute of Computer Applications and Management
A-4, Paschim Vihar, Rohtak Road, New Delhi – 110063 (INDIA)
Tel./Fax: 91 – 11 – 25275055 E-Mail: bijit@bvicam.ac.in
Visit us at www.bvicam.ac.in

Editorial

It is a matter of both honor and pleasure for us to put forth the inaugural issue of BIJIT; the BVICAM's International Journal of Information Technology. This first issue of the journal presents a compilation of fourteen papers that span a broad variety of research topics in various emerging areas of Information Technology and Computer Science. Some application oriented papers, having novelty in application, have also been included in this issue, hoping that usage of these would facilitate the overall economic growth. This issue is the first visible step in realizing our vision "*to achieve a standard comparable to the best in the field and finally become a symbol of quality*".

Our panel of expert referees posses a sound academic background and have a rich publication record in various prestigious journals representing Universities, Research Laboratories and other institutions of repute, which, we intend to further augment from time to time. Finalizing the constitution of the panel was a painstaking process, but it helped us to ensure that the best of the received manuscripts are showcased and that too after undergoing multiple review cycles, as required.

The fourteen papers that were finally published were chosen out of more than sixty papers that we received from all over the world for this inaugural issue. We understand that the launch of inaugural issue of this Journal was delayed from our planned schedule, but we also hope that you concur with us in the fact that quality review is a time taking process and is further delayed if the reviewers are senior researchers in their respective fields and hence, are hard pressed for time.

We wish to express our sincere gratitude to our panel of experts in steering the submitted manuscripts through multiple cycles of review and bringing out the best from the contributing authors. We thank our esteemed authors for having shown confidence in BIJIT and considering it a platform to showcase and share their original research work. We would also wish to thank the authors whose papers were not published in this inaugural issue of the Journal, probably because of the minor shortcomings. However, we would like to encourage them to actively contribute for the forthcoming issues.

The undertaken Quality Assurance Process involved a series of well defined activities that, we hope, went a long way in ensuring the quality of the publication. Still, there is always a scope for improvement, and so we request the contributors and readers to kindly mail us their criticism, suggestions and feedback at bijit@bvicam.ac.in and help us in further enhancing the quality of forthcoming issues.

Editors

CONTENTS

Fuzzy Logic Based Revised Defect Rating for Software Lifecycle Performance Prediction Using GMR

A. K. Verma[1], **Anil R**[2] and **Dr. Om Prakash Jain**[3]

Abstract - Software service organizations have adopted various software engineering process models and are practicing it earnestly. Even though this has helped the organizations to improve the quality and the profit margins; there exists a need to compare different groups within it so as to concentrate on the weaker sections. In this paper, the authors propose a revised model for defect rating that can be used for calculating group maturity within the organization. Fuzzy logic approach is used for the proposed model considering the linguistic or imprecise nature of the software measurements. The output of this model can be used as one of the parameter for predicting different software parameters within the software lifecycle.

.

Index Terms - Defect rating, Fuzzy logic, Historical data.

1. INTRODUCTION

Today, there exists many software reliability models [1],[2], [3], [4], [5], [6], [7], [8], [9] which predicts the defect density early in the life cycle. However, these models use the current trend of the defects for prediction. Most of these models are developed using some large software projects where the size of the source code is the range of many hundred thousand lines of codes. It is easy to develop and apply these models on large software projects because of the significant duration and effort spent. However, for industrial software projects that are of size less than a hundred thousand lines of code and being executed in less than six months, it is difficult to use these models for early prediction of the software defect density. In these projects, the average duration of testing may last only a couple of weeks.

Historical information from the past projects also needs to be used for the prediction of the defect density of the new projects [10]. The people and the maturity of the organization are playing an important role in the quality of the software being developed. One can not ignore these facts while predicting the quality of the software under development. In this paper, the authors propose a model that uses the historical information from the past projects and gives a rating for the present project which can be used along with other project parameters to predict the defect density of the project. Fuzzy logic approach is used for the developing the proposed model considering the advantages of fuzzy logic in converting the experts knowledge into fuzzy rules. The paper is organized in the following way. Section II introduces the concept of Fuzzy logic. Section III talks about the problem overview, Section IV talks about the parameters used for the model and the method of calculations, Section V talks about the proposed model using Fuzzy systems, section VI discuses about the application of the proposed model on industrial data and section VII concludes the paper along with future work.

2. FUZZY LOGIC

Fuzzy logic is invented by Zadeh in 1965. [11] [12]. It is being used in many important investigations since its invention. This concept provides a natural way of dealing with problems where the main source for impreciseness is the absence of crisply defined criteria. In fuzzy approach, the concerned phenomenon in the system is controlled by linguistic uncertainties. A typical fuzzy system consists of a fuzzifier, fuzzy engine and a defuzzifier. Due to the simplicity associated with it, Mamdani method is the most commonly used fuzzy interference engine even though there exists many other approaches [13], [14]. A sequence of fuzzy interface rules determines internal structure of the fuzzy engines. A typical fuzzy system consists of four steps.

1. Using membership functions, an input value is translated into linguistic terms. How much a given numerical input, which is under consideration, fits into the linguistic terms, is decided by the membership function.
2. Fuzzy rules are evolved by considering the different permissible combinations of input and output membership functions. The rules are defined with the use of experts' knowledge in the field under consideration.
3. The derived rules are applied to the membership functions and the aggregation of the outputs of the all rules takes place. This is performed by the fuzzy interference engine which maps the input membership function and the output membership function using the defined fuzzy rules.
4. Converting the resultant fuzzy output into a crisp number which is called as defuzzification.

3. PROBLEM OVERVIEW

Quality of the software being developed in an organization depends not only on the present project conditions, but also on the past performance of the group which develops the software. Considering this, a rating based on the historical data was developed using fuzzy logic technique. Group maturity rating (GMR) [15] is defined for predicting the software performance of a group with in a typical software organization of high maturity. This rating uses five parameters such as, schedule variance, effort variance, customer satisfaction index, process compliance index and defect rating. Since the parameters used for arriving at the model are either linguistic or data is uncertain or vague, fuzzy logic approach is considered as the best approach. Group maturity rating is being used as one of the environmental parameter apart from the project metrics for better prediction using Fuzzy-neuro approach.

[1] *Indian Institute of Technology, Mumbai*
[2,3] *L&T Infotech, Mumbai*

The defect rating used in the first version of GMR consists of two parameters, defect density and residual defect density and was defined using fuzzy logic approach. Even though, this model gives a good rating on the maturity of the groups under consideration, the defect rating can be refined by incorporating the review effectiveness as the third parameter, considering the fact that quality of the software under consideration depends of the effectiveness of the review which is being carried out. Also in some cases, the relative error with the existing model is on the higher side that can be reduced.

4. PARAMETERS UNDER CONSIDERATION

a. Defect Density

Defect density is one of the important metrics of software organizations and gives a picture of the quality of the projects of that organization. Defect density is defined as the defects per unit size of the software entity being measured. Low value of defect density is better, however, the same needs to be investigated, since ineffective review and testing also leads to low defect density. Defect density can be correlated with many parameters like the project management practices and processes followed by the project team, the technical knowledge of the organization, and on the competency of the people. Due to these factors, the historical information about the defect density of projects will always help the organization to decide on the time required for review and testing and stoppage rules of testing. Even though the defects found out during the review and testing are resolved before shipping, it takes a significant effort and time from the project. This will directly affect the profit of the organization. The membership functions for defect density are decided using the expert opinion and the historical baseline metrics. Trapezoidal membership functions are considered for defect density. The membership functions of defect density are decided as "Excellent", "Very good", "Good" and "Poor". The elements of the fuzzy sets are determined using the historical baseline mean and the control limits. Table I illustrates the formulae used to find out the membership values of defect density.

Membership function	Membership values
Very good	$0, \mu\text{-}\frac{9\sigma}{2}, \mu\text{-}\frac{7\sigma}{2}, \mu\text{-}\frac{5\sigma}{2}$
Good	$\mu\text{-}\frac{7\sigma}{2}, \mu\text{-}\frac{5\sigma}{2}, \mu-\frac{3\sigma}{2}, \mu-\frac{\sigma}{2}$
Poor	$\mu-\frac{3\sigma}{2}, \mu-\frac{\sigma}{2}, \mu+\frac{\sigma}{2}, \mu+\frac{3\sigma}{2}$
Very poor	$\mu, \mu+\sigma, \mu+2\sigma, \infty$

Table 1: Membership Values for Defect Density

b. Residual Defect density

Residual defect density shows the quality of the projects delivered by an organization and this is also one of the important defect metrics for an organization. Residual defect density (RDD) is the measure of the unresolved defects after release of the software entity per unit size. This number indicates the number of defects passed on to the customers after completing the in-house testing. RDD plays a crucial role in the customer satisfaction since it directly affects the customer whereas; DD defines in the quality of the in-house development. The membership functions for residual defect density are decided using the expert opinion and the historical baseline metrics. Trapezoidal membership functions are also considered for residual defect density. The membership functions of residual defect density are decided as "Excellent", "Very good", "Good" and "Poor". The elements of the fuzzy sets are determined using the historical baseline mean and the control limits. Table II illustrates the formulae used to find out the membership values of residual defect density.

Membership function	Membership values
Very good	$0, 0, \mu\text{-}\frac{3\sigma}{2}, \mu-\sigma$
Good	$\mu\text{-}\frac{3\sigma}{2}, \mu-\sigma, \mu+\frac{3\sigma}{4}, \mu+\frac{5\sigma}{4}$
Poor	$\mu+\frac{3\sigma}{4}, \mu+\sigma, \mu+\frac{13\sigma}{4}, \mu+\frac{15\sigma}{4}$
Very poor	$\mu+3\sigma, \mu+\frac{7\sigma}{2}, \mu+\frac{9\sigma}{2}, \infty$

Table 2: Membership Values for Residual Defect Density

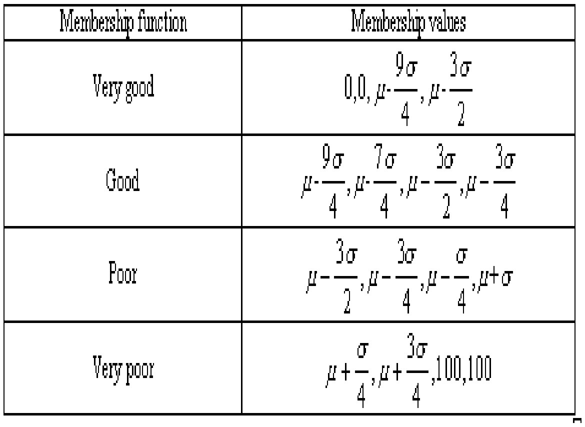
c. Review Effectiveness

During software development, there exist a lot of opportunities for errors. Even though, in ideal conditions, one expects no defects are injected during the development process, the same is an impossible target. In this scenario, the best possible method is to remove the maximum possible error injected as soon as possible. The first possible chance for finding out the errors while developing software is the review process.

Review effectiveness (RE) is the measure of the efficiency of the review process. It is the ratio of total defects found during reviews to the total no of defects found during the entire life cycle. This can be expressed as,

$$\text{RE}=\frac{\text{Number of defects found during review}}{\text{Number of defects found during lifecycle}}\times 100\%$$

The membership functions for review effectiveness also are decided using the expert opinion and the historical baseline metrics. For this parameter also, Trapezoidal membership functions are considered. The membership functions of residual defect density are decided as "Very Poor", "Poor", "Good" and "Very good". The elements of the fuzzy sets are determined using the



Membership function	Membership values
Very good	$0,0,\mu-\frac{9\sigma}{4},\mu-\frac{3\sigma}{2}$
Good	$\mu-\frac{9\sigma}{4},\mu-\frac{7\sigma}{4},\mu-\frac{3\sigma}{2},\mu-\frac{3\sigma}{4}$
Poor	$\mu-\frac{3\sigma}{2},\mu-\frac{3\sigma}{4},\mu-\frac{\sigma}{4},\mu+\sigma$
Very poor	$\mu+\frac{\sigma}{4},\mu+\frac{3\sigma}{4},100,100$

Table 3: Membership Values for Review Effectiveness

historical baseline mean and the control limits. Table III illustrates the formulae used to find out the membership values of review effectiveness.

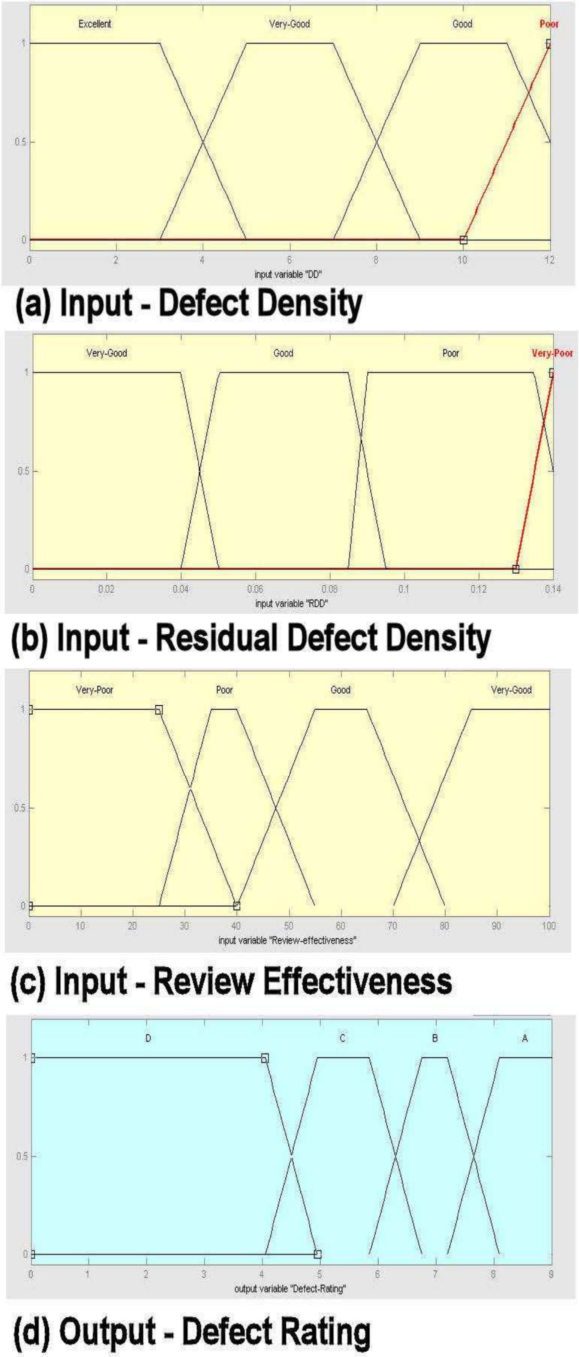
The output elements are selected as rating "A", "B", "C" and "D", where "A" is the best rating and "D" is the worst rating. These are chosen carefully with the help of experts in the field and converted into defect rules as stated in the next session.

5. DEFECT RATING

There exists a unique relationship between the parameters mentioned in the previous section. Considering this DD, RDD and RE are to be treated together. Low DD and low RDD is the best. When RDD is more and DD is less, it implies to the ineffective in-house testing and review. Here the influence of review effectiveness comes into picture. An effective review will definitely helps the defect densities to come down, but may not be in a linear scale. Considering these, a new parameter called Defect rating (DR) is developed using the different combinations of DD, RDD and RE. This will help the organization to know the health of the project. It also avoids the problem of comparing projects in different technologies since DD and RDD are correlated to the technology and review effectiveness is independent of technology.

A fuzzy logic model was created for defect rating. Sixty four different rules were created based on the input – output combination and fed to the fuzzy engine. Some of the example

- Rule 1: if (DD is **Poor**) and (RDD is **Very Poor**) and (RE is **Very Poor**)Then (Defect rating is **B**)
- Rule 11: if (DD is **Poor**) and (RDD is **Good**) and (RE is **Good**)Then (Defect rating is **C**)
- Rule 33: if (DD is **Very Good**) and (RDD is **Very Poor**) and (RE is **Very Poor**)Then (Defect rating is **D**)
- Rule 22: if (DD is **Good**) and (RDD is **Poor**) and (RE is **Poor**)Then (Defect rating is **C**)
- Rule 36: if (DD is **Very Good**) and (RDD is **Very Poor**) and (RE is **Very Good**)Then (Defect rating is **B**)
- Rule 48: if (DD is **Very Good**) and (RDD is **Very Good**) and (RE is **Very Good**)Then (Defect rating is **A**)
- Rule 52: if (DD is **Excellen**t) and (RDD is **Very Poor**) and (RE is **Very Good**)Then (Defect rating is **B**)
- Rule 61: if (DD is **Excellent**) and (RDD is **Very Good**) and (RE is **Very Poor**)Then (Defect rating is **B**)



(a) Input - Defect Density

(b) Input - Residual Defect Density

(c) Input - Review Effectiveness

(d) Output - Defect Rating

Figure 1: Membership Functions for Inputs and Output of Defect Rating

Mamdani method is used as the fuzzy interference engine. Defuzzified crisp output it taken as the input to the defect rating. Fig. 1 illustrates the mapping of inputs of the fuzzy logic into appropriate membership functions. The rules are created using the fuzzy system editor contained in the Fuzzy Logic Toolbox of Matlab 7.0. Control surface of Defect rating based on fuzzy rules is illustrated in Fig. 2 and Fig. 3. The fuzzy inference diagram in Fig. 4 displays all parts of the fuzzy inference process from inputs to outputs. Each row of plots corresponds to one rule, and each column of plots corresponds to either an input variable or an input variable. One can use the fuzzy inference diagram to change the inputs and to find out the corresponding outputs.

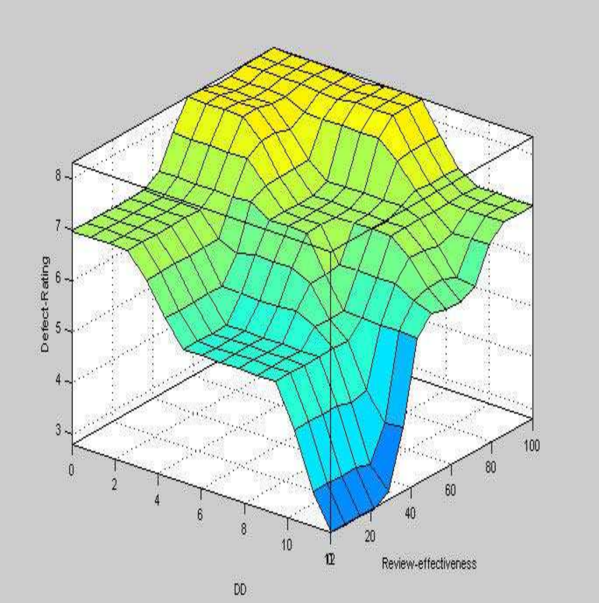


Figure 2: - Control surface for Defect rating fuzzy logic application – DD Vs RE

6. CASE STUDY

In order to validate the revised model of rating, new model is checked with the same set of industrial project data. The case study was employed with the data from six different groups from a typical software organization. The data set consists of data from 140 projects in the recent one year, which is filtered from a larger set of data to get a range of output. Outliers, which are the abnormal project data with large noise, are removed from the selected set of project data to arrive at best results. The project data is pre processed and removed five projects from the original database. The database is divided into three sets, based on the period of execution of these projects. Defect rating was calculated separately using the industrial data and the crisp output arrived from the defect rating model is fed as input to the Group maturity model. The

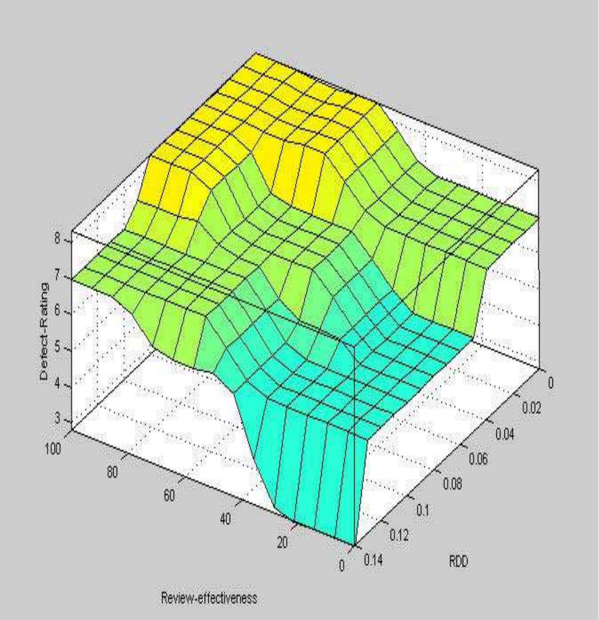
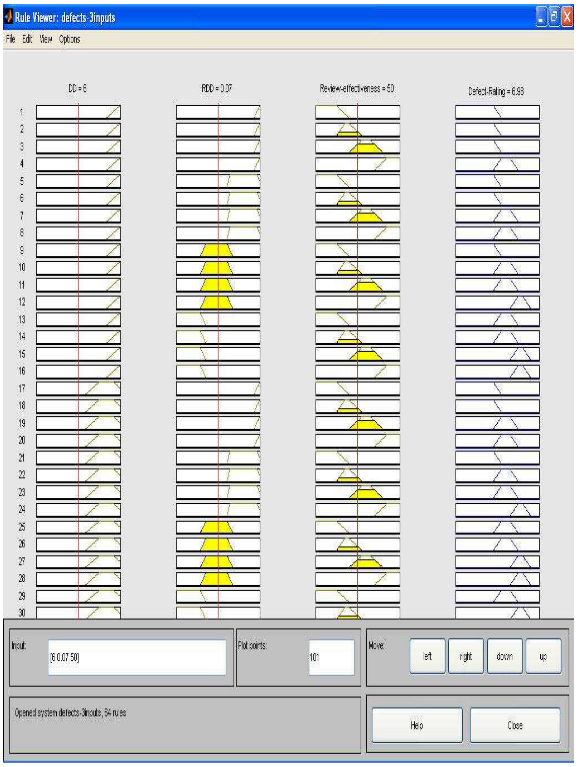


Figure 3: Control surface for Defect rating fuzzy logic application – RDD Vs RE



DD = 6

RDD = 0.07

Review-effectiveness = 50

Defect-Rating = 6.98

Input: [6 0.07 50]

Plot points: 101

Move:

Opened system defects-3inputs, 64 rules

Figure 4: Fuzzy inference diagram for Defect rating

output of the revised GMR was compared with the output of the earlier GMR and the same is produced in the table.

(a) Quarter 1

Group	Trad. rating	GMR		GMR Revised	
		Rating	MRE in %	Rating	MRE in %
Group 1	66.67	52.79	20.82	70.00	5.00
Group 2	66.67	70.00	5.00	73.98	10.97
Group 3	46.67	32.83	29.65	32.83	29.65
Group 4	40.00	32.34	19.14	32.34	19.14
Group 5	66.67	70.00	5.00	70.00	5.00
Group 6	53.33	46.33	13.13	46.33	13.14

(b) Quarter 2

Group	Trad. rating	GMR		GMR Revised	
		Rating	MRE in %	Rating	MRE in %
Group 1	60.00	43.02	28.30	70.00	16.67
Group 2	60.00	70.00	16.67	70.00	16.67
Group 3	33.33	31.63	5.10	31.64	5.09
Group 4	46.67	41.90	10.22	41.90	10.22
Group 5	66.67	53.05	20.43	53.01	20.49
Group 6	46.67	44.95	3.67	44.95	3.69

(c) Quarter 3

Group	Trad. rating	GMR		GMR Revised	
		Rating	MRE in %	Rating	MRE in %
Group 1	80.00	71.95	10.06	85.00	6.25
Group 2	60.00	70.00	16.67	70.00	16.67
Group 3	40.00	50.00	25.00	50.00	25.00
Group 4	60.00	55.96	6.74	55.97	6.71
Group 5	53.33	44.90	15.82	44.90	15.82
Group 6	26.67	31.82	19.31	31.81	19.28

Table 4: Comparison of Models

a. Evaluation Criteria

The criterion, Magnitude of relative error (MRE) is employed to asses and compare the performance of the model with respect to the existing model. It can be defined as

$$MRE = \frac{|\text{Excisting Rating-Group Maturity Rating}|}{\text{Excisting Rating}}$$

MRE value is calculated for each group i whose rating is to be determined.

To find out the mean error of the model, mean magnitude of the relative error is also determined, which can be calculated as

$$\mathrm{MMRE} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} MRE_i$$

The result of the evaluation is shown in the table IV. The mean magnitude of the relative error (MMRE) for the entire data set consisting of the data from all three quarters is reduced to 13.64% from 15.04% which was reported by the earlier model. Considering the vagueness and uncertain data and linguistic parameters, this error is well within the acceptable limit and the revised defect rating is performing better than the previous model.

7. CONCLUSION AND FUTURE WORK

In this paper, a revised model is proposed for rating the different groups within an organization based on the defect density, residual defect density and the review effectiveness of the historical projects. A fuzzy logic approach is used for defining the model. The model is used for calculating the group maturity rating for a typical software organization of high maturity. The model is then compared with the existing model and the results were discussed. It is observed that while applying the revised model, the man magnitude of the relative error is reduced to 13.64% from 15.04% reported by the previous model.

This paper offers some instances based on the current research into the aspect of using the historical data for predicting the various parameters of the software project throughout the development life cycle. Defect rating will be used as one of the environmental parameter apart from the project metrics for better prediction of software projects using fuzzy-neuro approach.

REFERENCES

[1] A. L. Goel and K. Okumoto, Time-dependent fault-detection rate model for software and other performance measures, IEEE Transactions on Reliability, 28, 1979, p 206-211

[2] H. Pham, A software cost model with imperfect debugging, random life cycle and penalty cost, International Journal of Systems Science, vol. 27, Number 5, 1996, p 455-463

[3] M. Ohba and S. Yamada, S-shaped software reliability growth models, Proc. 4th Int. Conf. Reliability and Maintainability, 1984, pp 430-436

[4] X. Teng and H. Pham,"A software cost model for quantifying the gain with considerations of random field environments", IEEE Transactions on Computers, vol 53, no. 3, 2004

[5] H. Pham, L. Nordmann, and X. Zhang, A General imperfect software debugging model with s-shaped fault detection rate, IEEE Transactions on Reliability, vol. 48, no. 2, 1999, p 169-175

[6] H. Pham and X. Zhang, An NHPP software reliability models and its comparison, International Journal of Reliability, Quality and Safety Engineering, vol.4, no. 3, 1997, p 269-282

[7] S. Yamada, M. Ohba, and S. Osaki, S-shaped reliability growth modeling for software fault detection, IEEE Transactions on Reliability, 12,1983, p 475-484

[8] S. Yamada and S. Osaki,"Software reliability growth modeling: models and applications", IEEE Transactions on Software Engineering, 11, 1985, p 1431-1437

[9] S. Yamada, K. Tokuno, and S. Osaki, Imperfect debugging models with fault introduction rate for software reliability assessment, International Journal of Systems Science, vol. 23, no. 12, 1992

[10] Business Unit Rating Ver. 5.0, Internal document, Larsen and Toubro Infotech limited, 2007.

[11] Lotfi A Zadeh, Fuzzy sets, Information and control, vol 8, pp 338-353, 1965.

[12] Lotfi A Zadeh, The concept of a linguistic variable and its application to approximate reasoning–I, Information Sciences, Volume 8, Issue 3, pp 199-249, 1975

[13] George K Klier and Tina A Folger, Fuzzy sets, Uncertainity and Information, Prentice Hall, 1988.

[14] Puyin Liu and Hongxing Li, Fuzzy Neural Network theory and Application, World Scientific, 2004.

[15] Ajit Kumar Verma, Anil R and Om Prakash Jain, "Fuzzy logic based Group Maturity Rating for Software Performance Prediction" International Journal of Automation and Computing, volume 4, issue 4, 406-412, October 2007

BIJIT - BVICAM's International Journal of Information Technology
Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi.

Genetic Algorithm: A Versatile Optimization Tool

Prof. Swati V. Chande[1] and **Dr. Madhavi Sinha**[2]

Abstract – Genetic Algorithms are a powerful search technique based on the mechanics of natural selection and natural genetics that are used successfully to solve problems in many different disciplines.

The good robustness of these algorithms on problems of high complexity has led to an increasing number of applications in the fields of artificial intelligence, numeric and combinatorial optimization, business, management, medicine, computer science, engineering etc.

In this paper we present a cross section of current genetic algorithm applications from diverse fields and lay a special emphasis on use of genetic algorithms in one of the most important optimization problems in computer science, database query optimization.

Index Terms - Genetic Algorithms, Query Optimization.

1. INTRODUCTION

'Genetic Algorithms' is one of the most useful, general-purpose problem-solving techniques available to developers. It has been used to solve a wide range of problems such as optimization, data mining, games, emergent behavior in biological communities etc.

Like other computational systems inspired by natural systems, Genetic Algorithms have been used in two ways, as techniques for solving technological problems, and as simplified scientific models that can answer questions about nature. [1]

In this paper we focus on the applications of Genetic Algorithms in problem solving. This paper is organized as follows: section 2 contains introductory material providing some general working principles of Genetic Algorithms, section 3 offers a view of the use of Genetic Algorithms to some real world problems, section 4 deals with applications of Genetic Algorithms to database query optimization and finally section 5 provides some concluding remarks and summary of the survey.

2. GENETIC ALGORITHMS

Genetic Algorithms, invented by John Holland, is an ABSTRACTion of biological evolution and is thus a method for moving from one population of chromosomes (strings of bits) to a new population by using a kind of natural selection together with the genetics inspired operators of recombination, mutation and inversion. Each chromosome consists of genes (bits) which are instances of allele (1 or 0) [1].

A Genetic Algorithm functions by generating a large set of possible solutions to a given problem. It then evaluates each of those solutions, and decides on a "fitness level" for each solution set. These solutions then breed new solutions. The parent solutions that were more "fit" are more likely to reproduce, while those that were less "fit" are more unlikely to do so. In essence, solutions are evolved over time [2]. This way there is evolution of the search space scope to a point where a solution can be found.

Figures 1 and 2 show the generic pseudocode and flow diagram of the complete genetic process respectively.

The steps of a general Genetic Algorithm are,

2.1 Representation:

An initial population is created from a random selection of solutions (which are analogous to chromosomes). It involves the representation of an individual (a possible solution or decision or hypothesis) in the form of its genetic structure (a data structure depicting a string of genes called chromosomes). At each point of the search process, a generation of individuals is maintained.

The initial population ideally has diverse individuals. This is necessary because the individuals learn from each other. Lack of diversity in population leads to sub-optimal solutions. The initial diversity may be arranged by uniformly random, grid initialization, non-clustering or local optimization methods [3].

2.2 Evaluation:

A value for fitness is assigned to each solution (chromosome) depending on how close it actually is to solving the problem (thus arriving to the answer of the desired problem). The fitness function is a measure of the objective to be obtained (maximum or minimum values). Fitness function is optimized using the genetic process [4] and evaluates each solution to decide whether it will contribute to the next generation of solutions [5].

Since it selects which individuals can reproduce and create the next generation of population, it is designed with care.

2.3 Selection:

Selection of individuals for the next generation to reproduce or to live on relies heavily on the evaluation function.

Those chromosomes with a higher fitness value are more likely to reproduce offspring (which can mutate after reproduction). The offspring is a product of the father and mother, whose composition consists of a combination of genes from them (this process is known as "crossing over").

After evaluating the fitness of the individuals, by applying the evaluation function, selection of 'fit' individuals for reproduction / recombination is done. The selection techniques that can be used are- deterministic selection, Proportional fitness, Tournament selection etc. each of the techniques has its own pros and cons and can be chosen depending on the problem and population at hand.

[1] *Principal (Computer Science), International School of Informatics and Management, Jaipur*

[2] *Reader, AIM and ACT, Banasthali Vidyapith*

2.4 Recombination:

Recombination or reproduction is as in biological systems, candidate solutions combine to produce offspring in each algorithmic iteration called a generation. From the generation of parents and children, the fittest survive to become candidate solutions in the next generation. Offsprings are produced by specific genetic operators, such as mutation and recombination.

Recombination is randomly picking one or more pairs of individuals as parents and randomly swapping segments (of genes) of the parents.

Solutions thus combine to form offspring for the next generation. Sometimes they pass on their worst information, but if recombination is done in combination with a forceful selection technique, then better solution results are obtained. Recombination may be performed using different methods such as 1-point recombination, n-point recombination and uniform recombination.

Mutation is the most basic way to alter a solution for the next generation. Operators from the local search techniques may be used to slightly twiddle with the solution and introduce new, random information.

It is thus brought about by randomly changing one or more digits (genes) in the string (chromosomes) representing an individual. In binary coding this may simply mean changing a 1 to a 0 and vice versa [6].

Elitism: There is a chance that the best chromosome may be lost when a new population is created by recombination and mutation. The best chromosome(s) may hence be copied to the new population. The rest is done in a classical way. This can rapidly increase the performance of the genetic algorithm, because it prevents the loss of the best-found solutions [7].

If the new generation contains a solution that produces an output that is close enough or equal to the desired answer then the problem has been solved. If this is not the case, then the new generation will go through the same process as their parents did. This will continue until a solution is reached.

```
Create initial population
    Repeat
        Evaluate each individual's fitness
        Select best-ranking individuals to reproduce
        Mate pairs at random among selected individuals
        Apply recombination operator
        Apply mating operator
Until termination condition
```

Figure 1: Pseudocode for a generic Genetic Algorithm

3. APPLICATION OF GENETIC ALGORITHMS TO SOME REAL WORLD PROBLEMS

3.1. Nutritional Counselling:

Nutrition counseling an important part of lifestyle counseling systems.

3.1.1. Gaál et al describe the method used by MenuGene, an intelligent menu planner that generates personalized dietary weekly menu plans with the emphasis on the prevention of Cardiovascular Diseases. The task of weekly dietary menu planning is considered a multi-objective optimization problem which is solved by a multi-level genetic algorithm. The algorithm decomposes the search space to sub-spaces according to the structure and nutritious content of the menu plan. [9]

3.1.2. Alkhalifa A.Y., Niccolai M.J., Nowack W.J. have developed a component of an information system for selection of a nutritionally, culturally, economically and geographically appropriate diet using Genetic Algorithm. [10]

3.2. Stylometry:

Holmes and Richard have applied genetic algorithms to create a set of rules for determining authorship and then let the most useful, or fit rules survive. They combine stylometry, the science of measuring literary or linguistic style, usually to written language, and genetic algorithms to determine authorship. [11]

3.3. Parametric Design of aircraft:

Marle F.Bramlette and Eugene E.Bonehard have discussed optimizing aircraft design when the task is posed as that of optimizing a list of parameters. They used real number representation for Genetic Algorithms and generated a large number of initial population members and worked only with the best ones. [12]

3.4. Robot trajectory generation:

This application demonstrates the application of Genetic Algorithm techniques to the task of planning the path which a robot area in moving from one point to another. Yuval Davidor uses variable-length chromosomes in this solution, & devises some novel & interesting crossover operators. [13]

3.5. Strategy acquisition for simulated airplanes:

A genetic algorithm (SAMUAL) that learns techniques for maneuvering a simulated airplane in order to evade simulated missiles has been described by John J. Grefenshelte. The SAMUEL system tries to discover rules by which a slower but more maneuverable aircraft can evade a faster but less agile missile until the missile runs out of fuel. [14]

3.6. Redistricting:

For a fair and equitable congressional redistricting of Texas, Michael Larson has given a simplified model based on the genetic algorithm technique using the TSP approach. [15]

3.7. Problem solving and in- circuit emulators. (Embedded systems).

An in-circuit emulator is a hardware debugging tool that through its on-board, modified processor, lets one emulate a line processor or family of processors on a target system.

One feature of an ICE is the ability to provide a program clock, essentially the heartbeat of the hardware being designed and tested. The setting up of the clock is done using genetic algorithm. [16].

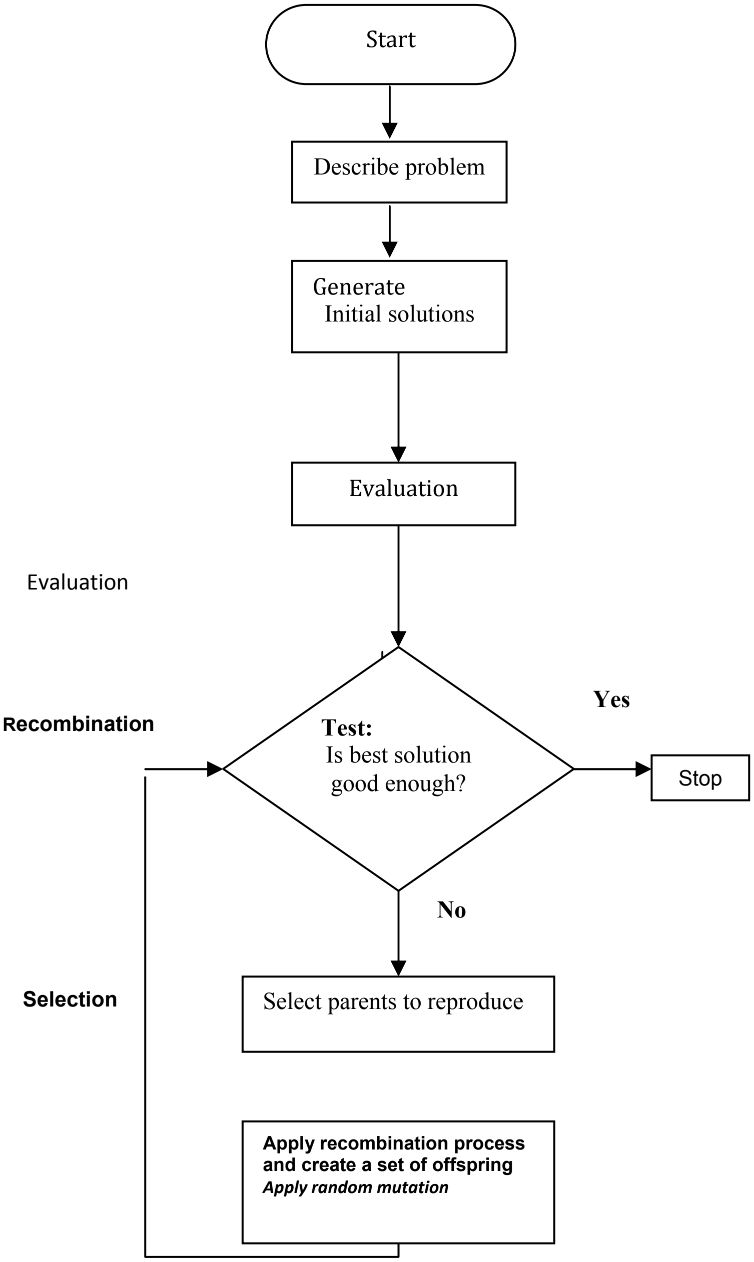


Figure 2: Flow diagram of the genetic algorithm process Adapted [16]

3.8. Acoustics:

Sato et al used genetic algorithms to design a concert hall with optimal acoustic properties, maximizing the sound quality for the audience, for the conductor, and for the musicians on stage. [17].

3.9. Aerospace engineering:

As telecommunications technology continues to improve, humans are increasingly dependent on Earth-orbiting satellites to perform many vital functions, and one of the problems engineers face is designing their orbital trajectories.

3.9.1. Williams, Crossley and Lang 2001 applied genetic algorithms to the task of spacing satellite orbits to minimize coverage blackouts. [19]

3.9.2 Lockheed Martin has used a genetic algorithm to evolve a series of maneuvers to shift a spacecraft from one orientation to another within 2% of the theoretical minimum time for such maneuvers. The evolved solution was 10% faster than a solution hand-crafted by an expert for the same problem. [19]

3.10. Bandwidth optimization in near video on demand system:

Kriti Priya Gupta has employed Genetic Algorithm technique for minimizing the average bandwidth requirement in near video on demand system. Near video on demand system enable customers requesting the same movie to be grouped together in batches & then the broadcasted to them, using multicasting in a simple transmission stream. [20]

3. 11. Medical

Genetic Algorithms can be used throughout the medical field. Genetic Algorithms can help develop treatment programs, optimize drug formulas, improve diagnostics, and much more.

Plasma X-ray Spectra Analysis: X-ray spectroscopic analysis is a powerful tool for plasma diagnostics. Golovkin et al use genetic algorithms to automatically analyze experimental X-ray line spectra and discuss a particular implementation of the genetic algorithm suitable for the problem. Since spectroscopic analysis may be computationally intensive, they also investigate the use of case injected genetic algorithms for quicker analysis of several similar (time resolved) spectra. [21]

3.12. Scheduling:

Genetic Algorithms can be used for numerous scheduling problems. Using a Genetic Algorithm for difficult scheduling problems enables relatively arbitrary constraints and objectives to be incorporated painlessly into a single optimization method.

3.12.1. Organizers of the Paralympics Games, 1992, used genetic algorithms to schedule events. [22]

3.12.2. School Timetabling:Geraldo Ribeiro Filho and Luiz Antonio Nogueira Lorena have given a constructive approach to the process of fixing a sequence of meetings between teachers and students in a prefixed period of time, satisfying a set of constraints of various types, known as school timetabling problem. Pairs of teachers and classes are used to form conflict-free clusters for each timeslot. Binary strings representing pairs are grouped based on dissimilarity measurement. Teacher preferences and the process of avoiding undesirable waiting times between classes are explicitly considered as additional objectives. [23]

3.13. Musical Composition

An approach of evolving form in musical composition is presented by Ayesh and Hugill. They use of genetic algorithms for this. The approach presented for genetic composition uses samples of musical ideas (one note or more) and not individual MIDI notes. The focus of the approach is on evolving musical form rather than attempting to compose musical sequences. The selection process is guided by the responses of the users within an interactive process. [24]

3.14. Finance:

3.14.1. Mahfoud and Mani 1996 used a genetic algorithm to predict the future performance of 1600 publicly traded stocks. Specifically, the Genetic Algorithm was tasked with forecasting the relative return of each stock, defined as that stock's return minus the average return of all 1600 stocks over the time period in question, 12 weeks (one calendar quarter) into the future. [25].

3.14.2. Naik 1996 reports that LBS Capital Management, an American firm headquartered in Florida, uses genetic algorithms to pick stocks for a pension fund it manages. [22]

3.14.3. Coale 1997 and Begley and Beals 1995 report that First Quadrant, an investment firm in California, uses Genetic Algorithms to make investment decisions for all of their financial services. [26, 27]

3.15. Identifying criminal suspects:

The "FacePrints" software, helps witnesses identify and describe criminal suspects. It uses a genetic algorithm that evolves pictures of faces based on databases of hundreds of individual features that can be combined in a vast number of ways. The program shows randomly generated face images to witnesses, who pick the ones that most resemble the person they saw; the selected faces are then mutated and bred together to generate new combinations of features, and the process repeats until an accurate portrait of the suspect's face emerges.[22]

3.16. Seeking Routes

Texas instrument is drawing on the skills that salmon use to find spawning grounds to produce a Genetic Algorithm that shipping companies can use to let packages "seek" their own best route to their destination.[22]

4. DATABASE QUERY OPTIMIZATION USING GENETIC ALGORITHMS

Optimization of queries can be done through two approaches, one consisting of algebraic manipulations or transformations and the other including strategies to take advantage of the storage of the relations. The algebra based optimization approach is to first represent each relational query as a relational algebra expression and then transform it to an equivalent but more efficient relational algebra expression. The transformation is guided by heuristic optimization rules [28]. The basic idea of the cost-estimation-based approach is - For each query, enumerate all possible execution plans. For each execution plan, estimate the cost of execution plan [28]. Finally choose the execution plan with the lowest estimated cost [29]. Enumerative strategies can lead to the best possible solution, but face a combinatorial explosion for complex queries (e.g., a join query with more than ten relations. Join operation is not only frequently used but also expensive [30].). In order to investigate larger spaces, randomized search strategies have been proposed [31] to improve a start solution until obtaining a local optimum. Examples of such strategies are simulated-annealing [Ioannidis 87] and iterative-improvement [Swami 88]. With the same objective, genetic search strategies [Goldberg 89] can be applied to query optimization, as a generalization of randomized ones [Eiben 90]. Randomized or genetic strategies do not guarantee that the best solution is obtained, but avoid the high cost of optimization. As an optimizer might face different query types (simple vs. complex) with different requirements (ad-hoc vs. repetitive), it should be easy to adapt the search strategy to the problem [32].

The major issue in query optimization is that, the search space is complicated and genetic algorithms are theoretically and empirically proven to provide robust search in complex spaces. These algorithms are computationally simple yet powerful in their search for improvement. They are not fundamentally limited by restrictive assumptions about search space [6]. The use of Genetic Algorithm approach in addressing the Query Optimization issue, therefore seems apt.

Genetic algorithms may be employed in obtaining an optimal solution for each of the two approaches. They may contribute towards the selection of an efficient relational algebra expression and may also find near-optimal execution plans through efficient cost estimation.

All query optimization algorithms primarily deal with joins. Most studies on the use of Genetic Algorithms in Query Optimization also thus focus on joins. Selection of appropriate index for query execution is also one of the major concerns and hence substantial research has also been done in the use of Genetic Algorithms for index selection.

Bennett et al [1991] have studied genetic algorithms for join query optimization. They have given a method for encoding arbitrary binary trees as chromosomes and describe several recombination operators for such chromosomes. Their performance results show that genetic algorithms can effectively identify high quality query execution plans, and the selected plans are in general comparable to or better than the current best--known method for query optimization particularly, the output quality and the time needed to produce such solutions [33].

Farshad Fotouhi and Carlos E. Galarce [1991] of Wayne State University Computer Science Department, have proposed using genetic algorithms to search for near-optimal indexing. They have used a single table and their gene is a binary vector with a position for each of the attributes in the table. A 1 means the column is indexed; 0 means it's not. There is no attempt to accept only genes with the primary key indexed. The idea is to let the genetic process find a solution without any help.

This same chromosome pattern can also be used to represent a type of query. A 1 in a "query chromosome" means that the corresponding column is to be returned; 0 means it's not. This correspondence makes it simple to simulate query runs. The payoff formula is based on hitting or missing an index in a query. The optimal score is to ask for only indexed columns, which makes sense because there's a chance that a non-indexed column would require a sequential search of the table. Fotouhi and Galarce ran a series of random queries with a known statistical distribution against the test database of one million rows. The genes with the highest scores were saved from ("survived") that test run and used to build the next test run ("generation"). The performance of the system was measured in terms of average query-response times. The system leveled out in about ten generations with a 5-bit chromosome, but took longer with a 10-bit chromosome.

The Fotouhi-Galarce experiment gave encouraging results but was based on a single table, a rare occurrence in the real world.

Celko [1993] extended the Fotouhi-Galarce experiment to work on multiple tables. He combined columns to make tables using normal forms built from Functional Dependencies. He used a sample database, identified the functional dependencies for the database and created a query chromosome structure having genes based on attributes. To show each possible 3NF schema, he built tables where functional dependencies for the database were the key. Once the tables were defined, queries were applied against the whole database schema.

A table chromosome was made up of a subset of the original functional dependencies. Two rules were to be obeyed by the tables. First, no combination that violates the 3NF condition was allowed; second, all columns must be present in some table in the schema;

The database schemas were made up of more than one table, and one or more tables were mutated at a time. The payoff function considered joins between tables, the number of tables accessed, etc.

Since the operations on the schema were complex than those used for indexes on a single table by Fotouhi and Galarce, tables had to combine or split. The goal was to have the smallest number of tables used in the queries to avoid the cost of joins. Once the tables were determined for the set of queries, the index genetic algorithm was applied to the tables [34].

Kratica, Ljubi´c and To¡si´c [2003] have proposed a genetic algorithm for solving the ISP (Index Selection Problem) i.e. the problem of minimizing the response time for a given database workload by a proper choice of indexes. Their Genetic Algorithm is based on binary encoding, data structures for the evaluation of the objective function, on the uniform crossover, and simple mutation. They have tested the algorithm on the class of challenging instances known from the literature and demonstrate that the results obtained indicate its efficiency and reliability [35].

Utesch's [1997] model attempts to find the solution of the QO problem similar to a traveling salesman problem (TSP). He has used Postgres Query Optimizer for the research. In this module possible query plans are encoded as integer strings, each string represents the join order from one relation of the query to the next. Parts of the GEQO module are adapted from D. Whitley's Genitor algorithm. Specific characteristics of the GEQO implementation in Postgres are, usage of a *steady state* Genetic Algorithm (replacement of the least fit individuals in a population, not whole-generational replacement) allows fast convergence towards improved query plans, this is essential for query handling with reasonable time; usage of *edge recombination crossover* which is especially suited to keep edge losses low for the solution of the TSP by means of a Genetic Algorithm; mutation as genetic operator is deprecated so that no repair mechanisms are needed to generate legal TSP tours. The GEQO module allows the Postgres query optimizer to support large join queries effectively through non-exhaustive search [36].

Lanzelottel and Patrick Valduriez [1991] have given a solution to the extensibility of the query optimizer search strategy. This solution is based on the object-oriented modeling of the query optimizer, where the search space and the search strategy are independently specified. It is illustrated by applying different search strategies including the genetic algorithm approach [32].

Steinbrunn, Moerkotte and Kemper [1997] have studied different algorithms that compute approximate solutions for optimizing join orders. They extensively scrutinized optimizers from the three classes, heuristic, randomized and genetic algorithms. From their study it turns out that randomized and genetic algorithms are well suited for optimizing join expressions. They generate solutions of high quality within a reasonable running time. The benefits of heuristic optimizers, namely the short running time, are outweighed by merely moderate optimization performance. This study concentrates on the generation of low-cost join nesting orders while ignoring the specifics of join computing.

Steinbrunn et al studied several algorithms for the optimization of join expressions and inferred that randomized and genetic algorithms are much better suited for join optimizations; although they require a longer running time, the results are far better.

For adequate solution space, they found that, with the exception of the star join graph, the bushy tree solution space is preferable in spite of the fact that "pipelining" (avoiding to write intermediate results to secondary memory) can be carried out mainly by left-deep processing trees.

Another consideration is the extensibility of randomized and genetic algorithms: both can be designed to optimize not merely pure join expressions, but complete relational queries. In addition, some of them (namely Iterative Improvement and genetic algorithms) can be easily modified to make use of parallel computer architectures [37].

The authors of this paper, motivated by the applicability of genetic algorithms in a wide range of problems and in optimization in particular, are working on the implementation of genetic algorithms to database query optimization. A Genetic Algorithm involving representation of joins as chromosomes, functions for evaluation of fitness and crossover and mutation operators is considered for minimizing the response time for a given database query.

5. CONCLUSION

Genetic Algorithms are good at taking larger, potentially huge search spaces and navigating them looking for optimal combinations of things and solutions which we might never be able to find. The use of genetic algorithms to solve large and often complex computational problems has given rise to many new applications in a variety of disciplines. They have discovered powerful, high quality solutions to difficult practical problems in a diverse variety of fields.

The few examples surveyed in this paper illustrate the diversity of approaches and point to some of the considerations that have proved important in making applications successful. The use of Genetic Algorithms, for example, in difficult scheduling problems, enables somewhat arbitrary constraints and objectives to be incorporated relatively easily into a single optimization method. With genetic algorithms, the focus lies on evolving forms, rather than on creating new solutions.

The choice of appropriate encoding scheme and fitness function determine the success of a genetic algorithm. Dembski[2002] has said that the 'fitness function guides an evolutionary algorithm into the target.[38]

In recent years, relational database systems have become the standard in a variety of commercial and scientific applications. This has augmented the demand for new, cost-effective optimization techniques for minimizing the response time for query. With genetic algorithms becoming a widely used and accepted method for very difficult optimization problems, their application to database query optimization seems apt. Genetic algorithms thus seem to offer an extremely effective, general purpose, means of dealing with both complexity and scale.

FUTURE SCOPE

Genetic Algorithms are of major significance to the development of the new generation of IT applications. The potential which they offer over existing techniques is enormous. They find application in biogenetics, computer science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields. And the list will continue to grow especially if Genetic Algorithms are combined with other optimization methods.

Current query optimization techniques are inadequate to support some of the emerging database applications. Genetic algorithms however, are ideally suited to the processing, classification and control of complex queries for very-large and varied data.

REFERENCES

[1] Melanie Mitchell, "An introduction to Genetic Algorithms", Prentice Hall of India, 2004

[2] Hsiung Sam, Matthews James, "An introduction to Genetic Algorithms", 2000 http://www.generation5.org/content/2000/ga.asp

[3] Singh Marjit M, "Genetic Algorithms: Inspired by Nature", Information Technology (it) Magazine, EFY, vol. 13, no.9, July 2004.

[4] Praveen Pathak, Michael Gordon, Weiguo Fan, "Effective Information Retrieval Using Genetic Algorithms Based Matching Functions Adaptation," hicss, vol. 02, no. 2, pp. 2011, February 2000. csdl.computer.org/comp/proceedings/hicss/2000/0493/02/04932011.pdf

[5] Luger G F, "Artificial Intelligence- structure and strategies for complex problem solving", 4th edition, Pearson Education, 2002.

[6] Goldberg David E, "Genetic Algorithms in search, optimization and machine learning", Pearson Education, 2003.

[7] Rajasekaran S, Pai G A Vijaylakshmi, "Neural Networks, Fuzzy Logic and Genetic Algorithms- synthesis and applications", Prentice Hall India, 2002.

[8] Turban E, Aronson J E, "Decision Support Systems, 6th edition, Pearson Education Asia, 2000.

[9] G. Gaál, I. Vassányi, and G. Kozmann, "Automated Planning of Weekly Dietary Menus for Personalized Nutrition Counseling", Proceeding 453, Artificial Intelligence and Applications 2/14/2005-2/16/2005 Innsbruck, Austria

[10] Alkhalifa, A.Y.; Niccolai, M.J.; Nowack, W.J., "Application of the genetic algorithm to nutritional counseling", Proceedings of the 1997 Sixteenth Southern, Biomedical Engineering Conference, 1997

[11] Holmes, David and Forsyth, Richard "The Federalist Revisited: New Directions in Authorship Attribution". Linguistic and Literary Computing, vol. 10, 1995, pp. 111–127

[12] Bramlette, M., Bouchard, E., 1991, "Genetic algorithms in parametric design of aircraft", Handbook of genetic algorithms, L. Davis, ed., Van Nostrand Reinhold, New York, pp. 109-123.

[13] Yuval Davidor "A Genetic Algorithm Applied To Robot Trajectory Generation", in Lawrence Davis, editor, Handbook of Genetic Algorithms, chapter 12, pages 144-165. Van Nostrand Reinhold, New York, New York, 1991.

[14] J. J. Grefenstette "Strategy acquisition with genetic algorithms", in L. Davis, editor, Handbook of Genetic Algorithms, pages 186--201. Van Nostrand Reinhold, 1991.

[15] Michael Larson, "Genetic Algorithms & optimal solutions", Developer 2.0, Dr. Dobb's journal, June 2004.

[16] Philip Joslin, "Genetic Algorithms & real world applications", Developer 2.0, Dr. Dobb's journal, June 2004.

[17] Sato, S., K. Otori, A. Takizawa, H. Sakai, Y. Ando and H. Kawamura. "Applying genetic algorithms to the optimum design of a concert hall." Journal of Sound and Vibration, vol.258, no.3, p. 517-526 (2002).

[18] Williams, Edwin, William Crossley and Thomas Lang. "Average and maximum revisit time trade studies for satellite constellations using a multiobjective genetic algorithm." Journal of the Astronautical Sciences, vol.49, no.3, p.385-400 (July-September 2001).

[19] Gibbs, W. Wayt. "Programming with primordial ooze." Scientific American, October 1996, p.48-50.

[20] Kriti Priya Gupta, "Genetic Algorithm approach for bandwidth optimization ", Synergy – ITS journal of I.T. & management, September 2005.

Continued on page no. 36

Data Hiding in JPEG Images

S.K.Muttoo[1] and **Sushil Kumar**[2]

1. INTRODUCTION

Steganography has been an important subject since people started communicating in writing. Steganography means hiding a secret message (the *embedded message*) within a larger one (*source co*ver) in such a way that an observer cannot detect the presence of contents of the hidden message. Today the growth in the information technology, especially in computer networks such as Internet, Mobile communication, and Digital Multimedia applications such as Digital camera, handset video etc. has opened new opportunities in scientific and commercial applications. But this progress has also led to many serious problems such as hacking, duplications and malevolent usage of digital information. Steganography finds its role in attempt to address these growing concerns. We know that, with the use of steganographic techniques, it is possible to hide information within digital audio, images and video files which is perceptually and statistically undetectable. The method of embedding secret message (which can be plain text, cipher text, or even images) is usually based on replacing bits of useless or unused data in the source cover (can be audio files, sound, text, Disk space, hidden partition, network packets, digital images, software, or circuitry). There are two common methods of embedding: *Spatial embeddin*g in which messages are inserted into the LSBs of image pixels, and *Transform embedding* in which a message is embedded by modifying frequency coefficients of the cover image (result is called the *stego-image*).Transform embedding methods are found to be in general more robust than the Spatial embedding methods which are susceptible to image-processing type of attacks. However with respect to steganography robustness is not a critical property but the perceptibility (i.e., whether the source cover is distorted by embedding information to a visually unacceptable level). There is another important issue of steganography, namely, capacity, i.e., how much information can be embedded relative to its perceptibility [5, 1].

We shall use digital images as the cover object in this paper in which we embed the hidden information. The challenge of using steganography in cover images is to hide as much data as possible with the least noticeable difference in the stego-image. Steganographic algorithms operate on basically three types of images: Raw images (i.e., bmp format), Palette based images (i.e., GIF images) and JPEG images. JPEG images are routinely used in Steganographic algorithms due to the most popular lossy image compression method. Usually it is found that an algorithm used to hide large amounts of information typically result in lower perceptibility (i.e., greater change to the image appearance) and a more robust algorithm result into lower embedding capacity. The JPEG image generation first decomposed the input image into a number of 8 x 8 blocks. Then DCT of each block are computed and the resultant DCT coefficient matrix is quantized using a standard quantization table. Finally the inverse DCT of quantized coefficient matrix are evaluated and the final JPEG image is obtained after rounding the values.

1.1. Jpeg-Jsteg

One of the well known embedding method of steganography based on Transform domain is Jpeg-Jsteg which embeds secret message (that is, in encoded form with help of Huffman codes) into LSB of the quantized DCT coefficients. There is one disadvantage of Jpeg-Jsteg that only few messages can be embedded in the cover-image. Also, Andreas Westfeld and Andreas Pfitzmann [11] noticed that steganographic systems that change LSBs sequentially cause distortionsdetectable by steganalysis methods. They observed that for a given image, the embedding of high-entropy data (often due to encryption) changed the histogram of color frequencies in a predictable way. J.Fridrich [3] has claimed that her method can potentially detect messages as short as any single bit change in a JPEG image.

Chang [2] has proposed a new Steganographic method to increase the message load in every block of the stego-image while retaining the stego-image quality. He has suggested a modified quantization table such that the secret message can be embedded in the middle-frequency part of the quantized DCT coefficients. Moreover, his method is as secured as the original Jpeg-Jsteg.

Neils Provos [7] has proposed another method to counter the statistical attack known as OutGuess. In the first pass, similar to Jsteg, OutGuess embeds message bits using a pseudo-random number generator to select DCT coefficients at random. After embedding, the image is processed again using a second pass. This time, corrections are made to the coefficients to make the stego-image histogram match the cover image histogram.

1.2. T-codes

We know that the best variable-length codes (VLC) are the Huffman codes. They are easy to construct for optimum efficiency if source statistics are known. But if they are used in serial communication, a loss of synchronization often results in a complex resynchronization process whose length and outcome are difficult to predict T-codes provide the solution to this problem.

T-codes are families of variable-length codes (VLC) that exhibit extraordinarily strong tendency towards self-synchronization. The concept of simple T-codes were given by M.R. Titchner [8]. In 1996 [9], Titchner proposed a novel recursive construction of T-codes known as the Generalized T-codes that retain the properties of self-synchronization.

[1]Reader, Department of Computer Science, University of Delhi, India

[2]Reader, Rajdhani College, University of Delhi, New Delhi, India

E-mail:[1] skmuttoo@cs.du.ac.in and azadsk2000@yahoo.co.in

Gavin R. Higgie [4] showed that in situation where codeword synchronization is important, the T-codes can be used instead of Huffman codes, giving excellent self-synchronizing properties without sacrificing coding efficiency. The main advantage of the T-Code is that they are self-synchronizing, so if some bits are lost or modified in a T-code encoded stream, the decoder will regain synchronization automatically. The best T-codes achieve self-synchronization within 1.5 characters following a lock loss. Thus, we can use T-codes in place of Huffman codes in the algorithms such as Jpeg-Jsteg. The advantage of this approach is the ability to send steganographic messages in lossy environment that are robust against detection or attack.

A modified robust steganographical method using T-codes is proposed by Muttoo and Sushil [6] and is compared with steganography methods based on Jpeg-Jsteg and Outguess techniques. They have shown that using of T-codes as source encoding in place of Huffman codes result into better PSNR values.

In this paper we propose T-codes for the encoding of original message and for the entropy encoding of compressed stego-image in place of Huffman codes. The proposed scheme takes advantage of the synchronizing ability of T-codes to increase the robustness of popularly used hiding techniques like Jpeg-Jsteg.

2. PROPOSED ALGORITHM

We have developed a novel steganographic method based on Jpeg-Jsteg, famous hiding-tool based on joint photographic expert group(JPEG).The embedding and extracting algorithms are summarized as under:

Embedding Algorithm 2.1:

```
Input: secret message, cover image
Procedure:
Step1.  Encode the message using the T-codes
Step2.  Divide the cover image into 8x8 blocks
Step3.  Calculate DCT coefficients for each block
Step4.  Quantize the coefficients
Step5.  while complete message not embedded do
          5.1  get next DCT coefficient
          5.2  if DCT ≠ 0 , DCT ≠1 and DCT ≠ -1 then
                 5.2.1  get next bit from message
                 5.2.2  replace DCT LSB with message bit
               end{if}
        end{While}
Step6.  De-quantize and take inverse DCT to obtain stego-image
End.
Output: Stego- image
```

Extracting Algorithm 2.2:

```
Input: Stego image
Procedure:
Step1.  Divide the stego image into 8x8 blocks
Step2.  Calculate DCT coefficients for each block
Step3.  Quantize the coefficients
Step4.  while secret message not completed do
          4.1  get next DCT coefficient
          4.2  if DCT ≠ 0 , DCT ≠1and DCT ≠ -1 then
                 Concatenate DCT LSB to secret message
               end{if}
        end{while}
Step5.  Decode secret message bits using the T-codes
end.
Output:  Secret message
```

3. RESULTS AND ANALYSIS

The proposed algorithm has been implemented on number of images. The measures such as message capacity (Codeword length) and PSNR values for the two gray-level cover images of size 128 x 128, namely, 8.tif (figure 3.1) and 0.tif (figure 3.2) and four 256 x 256 pixels, namely, Lena, Baboon, Peppers and tree (figures 3.3 to 3.6) obtained from the proposed algorithm are compared with the corresponding Jpeg-Jsteg method.

The figures 3.7 and 3.8 are stego-images obtained by Jpeg-Jsteg using Huffman and proposed T-codes used Jpeg-Jsteg method.

The results of PSNR values obtained from the existing Jpeg-Jsteg and proposed methos are summarized in two tables: Table 3.1 and table 3.2.

In Table 3.1 , we find that when the embedding capacity is increased, PSNR values of the proposed method are close to the PSNR values of the existing method, showing that the proposed method is as good as the existing method

		Jpeg-Jsteg (Huffman)	Jpeg-Jsteg (T-codes)
Image	Message Capacity	PSNR	PSNR
Lena	4382	37.77	37.69
Baboon	6026	36.49	36.40

		Jpeg-Jsteg (Huffman)	Jpeg-Jsteg (T-codes)
Peppers	4403	37.77	37.83
Tree	5554	36.78	36.70

Table 3.1

In Table 3.2, we notice that PSNR values of the proposed method are better than the existing method

		Jpeg-Jsteg (Huffman)		**Jpeg-Jsteg (T-codes)**
Image	Code word length	PSNR	Code word length	PSNR
8.tif	852	34.97	904	35.26
0.tif	650	34.47	681	35.81

Table 3.2

Test Images (Figure 3.1 to 3.6)

Fig. 3.1 (0.tif (16.3 KB))

Fig. 3.2 (8.tif (4.10 KB))

Fig. 3.3 (Lena (93.7 KB))

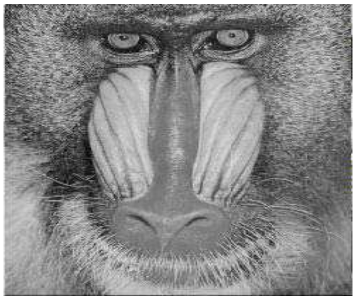


Fig. 3.4 (Baboon(97.0 KB))

Fig.3.5 (Peppers (89.6 KB))

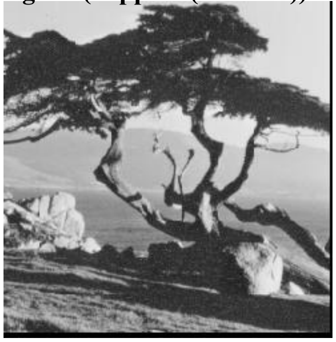


Fig. 3.6 (Tree (82.1 KB))

Stego- images (Figures 3.7 & 3.8):

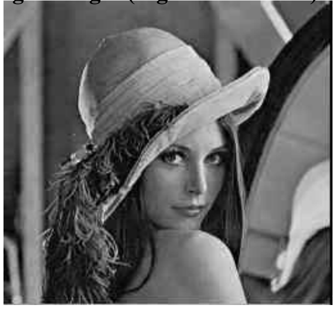


Fig. 3.7 (PSNR = 37.77 dB) Jpeg-Jsteg (Huffman)

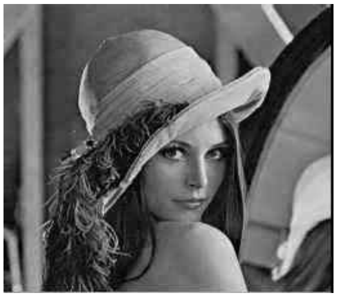


Fig. 3.8 (PSNR = 37.57 dB) Jpeg-Jsteg (T-codes)

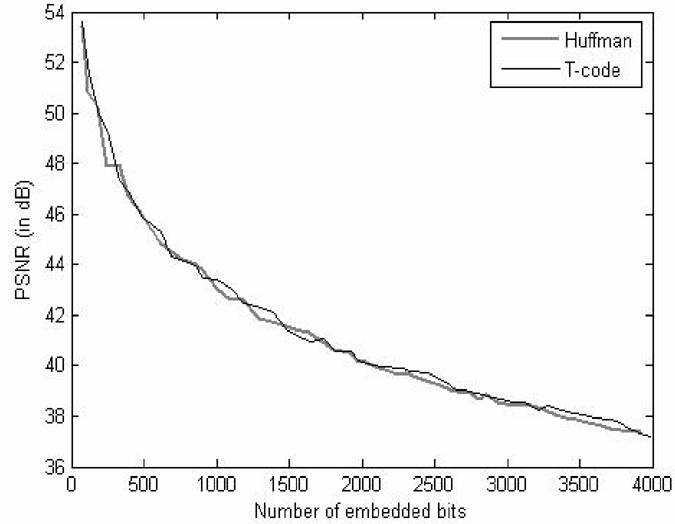


Figure 3.9: PSNR vs embedded bits

In the figure 3.9 we have shown the comparison of PSNR versus embedded message capacity by the original (Huffman based Jpeg-Jsteg) and the modified (T-code based Jpeg_Jsteg). The Dark lines are shown for T-code method where as light gray lines are for the Huffman method. We observe that the variation in PSNR values obtained with increasing values of embedded capacity is almost equal to the original method.

4. CONCLUSION

We observe from our experimental results that PSNR values of the proposed Jpeg-Jsteg algorithm based on T-codes for the different images are almost same as that of original algorithm based on Huffman codes, i.e., there is no change in the stego-image quality. Our method is secure in the way that even if the attacker detects (i.e., statistical attacks) and extracts the embedded message from the stego-image, he/she would not be able to recover the secret message without the encoded key. Moreover due to the inherent property of self-synchronizing of T-codes, our method is more robust as after the extraction process the recovered secret message is decoded and found to be without being much destroyed (for results one may refer to [12])

FUTURE WORK

Our approach has been to develop a Steganographic method that is perceptible and robust. It is known that JPEG approach can be statistically attacked even at one bit embedding. We are in the process of applying 'Best' T-codes as described by Ulrich Gunther [10] in place of simple T-codes to make the algorithm more secure from the attack such as filtering, cropping, noise etc. We are also applying this method to other data hiding techniques. This work will be presented in the form of a research paper in due course of time.

ACKNOWLEDGEMENT

The authors wish to thank their students Ram Kumar Karn, Rohit Choudhary and Kumar Sarth for their help in Matlab implementation.

REFERENCES

[1] Anderson, R.J. and F.A.P. Petitcolas, *" On the limits of steganography"*. IEEE J. Selected Areas in Commun., 16: 4, 1998.

[2] Chin-Chen Chang, Tung-Shou, and Lou-Zo Chung, *"A steganographic method based upon JPEG and quantization table modification"*, Information Sciences 141 ,123-138, 2002

[3] Fridrich J., Goljan M.& DU R, "Steganalysis Based on JPEG Compatibility', Special session on Theoretical and Practical Issues in Digital Watermarking and Data Hiding, SPIE Multimedia Systems and Applications IV, Denver, CO. August 20-24, 2001.

[4] G.R.Higgie, *"Analysis of the Families of Variable-Length Self-Synchronizing Codes called T-codes"*, Ph.D thesis, The University of Auckland, 1991.

[5] Johnson, N.F. and S. Jajodia,*"Exploring steganography: Seeing the unseen"*, IEEE Computer, 31: 26-34,1998.

[6] Muttoo S.K. and Sushil kumar ," Robust Steganography using T-codes", Proceeding of National Conference on Computing for nation development, IndiaCom 2007, Sponsored by AICTE, CSI , and IETE , Delhi, pp.221-223

[7] Niels Provos and Honeyman, *"Hide and Seek: an Introduction to Steganography"*, IEEE Security and Privacy, 32-43, May/June, 2003

[8] Titchener, M.R.,*"Technical note:Digital encoding by way of new T-codes"*, IEE Proc. E. Comput. Digit Tech., 1984, 131, (4),pp. 151-153.

[9] Titchener, M.R., *"Generalised T-codes: extended construction algorithm for self- synchronization codes"*. IEE Proc. Commun.,Vol. 143, No.3, 122-128, 1996

[10] Ulrich Gunther, "Robust Source Coding with Generalised T-codes", HhD Thesis, The University of Auckland, 1998. http;//www.tcs.Auckland.ac.nz/~ultivh/phd.pdf

[11] A. Westfeld and A. Pfitzmann, *"Attacks on steganogrphic systems"*, 3rd International Workshop on Information Hiding, 1999.

[12] Muttoo S.K. and Sushil kumar , "Image Steganography using Self-synchronizing variable codes ", International Conference on quality, reliability and Infocom technology, ICQRIT 2006.

Pattern Matching Based Technique to Solve Motif-Finding Problem

Pankaj Agarwal[1] and **Dr. S.A.M. Rizvi**[2]

Abstract - ***The problem of finding motifs from multiple molecular sequences is considered to be a difficult problem in molecular biology. In fact it is considered to be a Non-Deterministic Polynomial (NP)-complete problem and constant research is been carried out to solve the problem using some deterministic algorithm in polynomial time. Finding motifs from a set of DNA sequences is a critical step for understanding the gene regulatory network. This paper is an attempt to solve the motif problem using a deterministic method in polynomial time. The proposed method is not an exact algorithm but the probability of success is quite high by using it. Significance of the technique is its simplicity and time efficiency. The proposed technique is implemented as one of the modules in our general-purpose tool by the name "Sequence Comparison and Analysis Tool" for solving a number of sequence comparison problems encountered in the field of bioinformatics.***

1. INTRODUCTION

One important problem in bioinformatics is to understand how genes function in a gene regulatory network. Related to this is a sub problem of finding motifs for co-regulatory genes. A *gene* (protein coding gene) is a segment of DNA that codes for a specific protein. Genes seldom work alone. In most cases, genes cooperate to produce different proteins to provide particular functions. Understanding how the gene regulatory network works is important in molecular biology. In order to start the decoding process (*gene expression*), a molecule called *transcription factor* binds to a short region (*binding site*) preceding the gene. A *transcription factor* is a protein that regulates the activation of transcription in the eukaryotic DNA. Transcription factors localize the regions of promoter and enhancer sequence elements either through direct binding to DNA or through binding other DNA-bound proteins.

Transcription factor can bind to the binding sites of several genes to cause these genes to co-express. These binding sites have similar patterns called *motifs* [1]. Finding motifs from a set of DNA sequences is a critical step for understanding the gene regulatory network. In general by ``motifs", we refer to specific regions within larger DNA sequences that have some specific function. For example restriction sites are an example of a short sequence within a DNA molecule that has the function of being recognized by restriction enzymes. Motifs are generally short patterns (and usually but not always ungapped) and may be defined for DNA, RNA or Protein sequences. The discovery of motifs will allow the biologist to understand the varied and complex mechanism that regulates gene expression [2]. The objective of this paper is to devise a simple & effective methodology for determining motifs of any size from multiple molecular sequences.

2. RELATED WORK

Solving motif-finding problem has always been one of the key areas of interest for the researchers in the field of bioinformatics. Number of methods, algorithms and tools have come up in the recent years. Few of the common methods have been considered and discussed here.

CONSENSUS [2, 3] is a greedy algorithm requires no additional prior information other than the size of the desired motif. Generally, it works by extracting all possible subsequences of the correct length that are found in the sequences. Then it iteratively combines these subsequences together and calculates the positional weight matrix (PWM) for each set, keeping the best ones at each step.

Gibbs sampling [4] approach starts with a guess for where a motif is located in each input sequence and then uses those guesses to make more informed guesses. It chooses motif locations in a semi-random fashion, so it is not a greedy algorithm, but it is affected by where the initial guesses are located.

Expectation Maximization (EM) [5] is a term for a class of algorithms that estimates the values of some set of unknowns based on a set of parameters (the so-called "Expectation step"), and then uses those estimated values to refine the parameters (the "Maximization step"), over several iterations.

SP-STAR Combinatorial approach [6] was proposed by **Pevzner and Sze, 2000**. First, it chooses a suitable scoring function to access the goodness of a motif. Then, for each *l*-mer appearing in the sample, it finds the best instance in each sequence and collects these instances together to form an initial motif. It then employs a local improvement heuristic to improve each initial motif.

In recent times many new methods and algorithms have been proposed [7, 8, 9, 10]. A recent comparison of 13 current motif-Finding tools has been made available on the web page http://bio.cs.washington.edu/assessment

3. PROPOSED METHOD

This work concentrates only on *planted motif problem*, which is defined as:

Let $S = \{S_1, S_2 \ldots S_m\}$ be a set of sequences. For a length l pattern M, define $\delta(S_i\ ,M)$ to be the minimum number of substitutions between S_i and M. Define *score* $(M) = \sum \delta(S_i\ ,M)$. For example, suppose $S = \{S_{1,}\ S_{2,}\ S3,\ S4\}$, where

$S1$=TAGTACTAGGTCGGACTCGCGTCTTGCCGC
$S2$=CAAGGTCCGGCTCTCATATTCAACGGTTCG
$S3$=TACGCGCCAAAGGCGGGGCTCGCATCCGGC

[1] *Asst. Professor, Krishna Institute of Engineering and Technology, Department of Computer Science and Engineering, Ghaziabad, U.P.*

[2] *Head, Department of Computer Science, Jamia Millia Islamia Central University, New Delhi.*

*S*4=CCTCTGTGACGTCTCAGGTCGGGCTCTCAA
Here *M* = AGGTCGGGCTCGCAT.

Then we have $\delta(S_1, M) = 2$ as sequence S_1 and *M* differ in their respective positions at two places Similarly $\delta(S_2, M) = 2$, $\delta(S_3, M) = 2$ and $\delta(S_4, M) = 2$. Thus, *score* $(M) = 2 + 2 + 2 + 2 = 8$.

A set S of sequences each of length ‘n’ and two integers L & D with D<L<n acts as input to the proposed algorithm. The output is a pattern M such that every sequence in S contains a length L sub-string that can be transformed to M using at most D substitutions. For explanation of the algorithm, following set of four sequences have been considered

*S*1=TAGTACTAGGTCGGACTCGCGTCTTGCCGC
*S*2=CAAGGTCCGGCTCTCATATTCAACGGTTCG
*S*3=TACGCGCCAAAGGCGGGGCTCGCATCCGGC
*S*4=CCTCTGTGACGTCTCAGGTCGGGCTCTCAA

3.1 ALGORITHM

Step1: As a first step all the sub-strings corresponding to window size L from the sequence S_1 are stored. Sub-strings are obtained by considering a window of size L and shifting the window by one from left to right till all the sub-strings are collected in a table as depicted below in the Table 1.

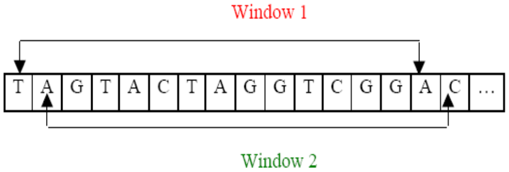


Figure 1: Two windows of size L=15 are shown.

For the considered sequence S_1 following sub-strings of window size L=15 can be obtained

Step 2: Now using the divide & conquer strategy FAT Tree is constructed for each of the obtained sub-strings corresponding to windows. Value for either height or level of the tree is also stored. A sample of FAT Tree is depicted below in the figure 2.

Step 3: Searching process begins here. For each of the obtained sub-sequence of window size L, it is searched in rest of the sequences. For a sample subsequence searching is carried out in the following manner:

a) Pattern at rood node is first searched in all the sequences (here S2, S3, S4). If found (exact pattern) it is stored along with the sequence number and calculated percentage of matched characters (here obviously 100%) in a table.

b) If the pattern is not found, counter value associated with the present sequence is incremented by one and then search is carried out starting with left node (here node associated with pattern ‘TAGTACT’). If pattern at left node is found, then we search for the remaining part of the sequence by following the link part (here link between the left and right node is depicted in the figure 2 by dotted arrows). It is to be remembered that remaining part of the pattern should be searched immediately after the position where its first part was matched in the sequence under consideration.

Window 1	TAGTACTAGGTCGGA
Window 2	AGTACTAGGTCGGAC
Window 3	GTACTAGGTCGGACT
Window 4	TACTAGGTCGGACTC
Window 5	ACTAGGTCGGACTCG
Window 6	CTAGGTCGGACTCGC
Window 7	TAGGTCGGACTCGCG
Window 8	AGGTCGGACTCGCGT
Window 9	GGTCGGACTCGCGTC
Window 10	GTCGGACTCGCGTCT
Window 11	TCGGACTCGCGTCTT
Window 12	CGGACTCGCGTCTTG
Window 13	GGACTCGCGTCTTGC
Window 14	GACTCGCGTCTTGCC
Window 15	ACTCGCGTCTTGCCG
Window 16	CTCGCGTCTTGCCGC

Table 1: A list all the sub-strings corresponding to window size 15 and window shift of 1

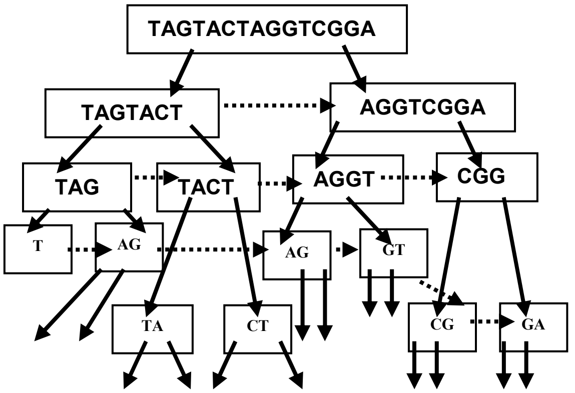


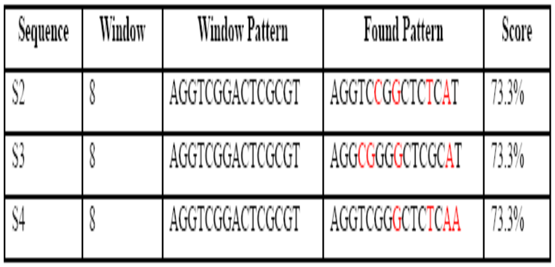
Figure 2: A FAT-Tree corresponding to pattern “TAGTACTAGGTCG”

c) Now if suppose pattern associated with right node is not found resulting in the increase of counter value further by one, then both the child nodes (here nodes with patterns ‘AGGT’ and ‘CGGA’) are exploited starting from left node.

d) While searching if at any instance of time counter value exceeds the value given as $2^{lvl}/2$ where ‘lvl’ refers to the level of the tree, then algorithm assumes that the considered sequence should be ignored from the process.

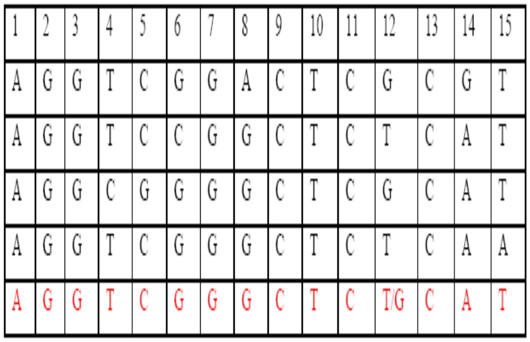
Step 4: A table is constructed with four values namely sequence number, window number, sub-sequence obtained from the first taken sequence, pattern found in the searched sequence and it’s associated score in percentage calculated as Score=(number of matched characters in the taken sub-sequence/L) *100 as depicted in table 2. Now table can be scanned to find the entries with maximum percentage for each set of considered sequences (here S1,S2, S3 & S4).

Step 5: All the maximal sets of patterns obtained are arranged in another table with separate rows for each pattern. Characters with maximum frequency at each position are then collected to give the final motif of length L as depicted in table 3



Sequence	Window	Window Pattern	Found Pattern	Score
S2	8	AGGTCGGACTCGCGT	AGGTCCGGCTCTCAT	73.3%
S3	8	AGGTCGGACTCGCGT	AGGCGGGGCTCGCAT	73.3%
S4	8	AGGTCGGACTCGCGT	AGGTCGGGCTCTCAA	73.3%

Table 2: The window patterns and associated found patterns with maximum score



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	G	G	T	C	G	G	A	C	T	C	G	C	G	T
A	G	G	T	C	C	G	G	C	T	C	T	C	A	T
A	G	G	C	G	G	G	G	C	T	C	G	C	A	T
A	G	G	T	C	G	G	G	C	T	C	T	C	A	A
A	G	G	T	C	G	G	G	C	T	C	T/G	C	A	T

Table 3: Characters with maximum frequency at each position

Thus the final motif can be represented as
M=AGGTCGGGCTCTCAT or
AGGTCGGGCTCGCAT

The described method is based on dynamic programming approach, which is a well-known technique to solve optimization problem. Here the problem of finding motifs from given molecular sequences have been considered as an optimization problem. Since every possible sub solution (sub sequences) is considered and a matching algorithm is used to determine the degree of match in each iteration where the scores of match are stored within a table followed by final scanning of the table to give the most optimal match; the chances of failure is quite low.

3.2 Time Complexity Analysis

The above-presented proposed method is carried out basically in five steps:

Extraction: All the sub-strings equivalent to window size L are extracted from one of the considered input sequence. This process will take at most O(n) worst-case time-complexity considering that the sequence has 'n' characters.

FAT-tree Construction: construction of the FAT-Tree corresponding to each of the extracted patterns in above step will take O(k.lg(L)) worst-case time complexity where k=number of extracted patterns each of length L with k<n.

Searching and Table construction: Searching each of the extracted patterns in all the remaining input sequences of length 'n' and 'm' being the number of such sequences will take O(k.m.n.lg(L)) time complexity.

Table Scanning: this should take atmost O(m.k)
Frequency Calculation: this step will take O(m.n)+O(n) complexity

The final time complexity can thus be given as
T(n)= O(n)+ O(k.lg(L))+ O(k.m.n.lg(L))+ O(m.k)+ O(m.n)+O(n)=O(k.m.n.lg(L)) which can be further given as $O(k.n^2.lg(L))$ provided n=m.
As already stated that motif-finding problem is considered as NP-Complete problem and solving the problem for a given small set sequences by using some deterministic algorithm in polynomial time is always significant.

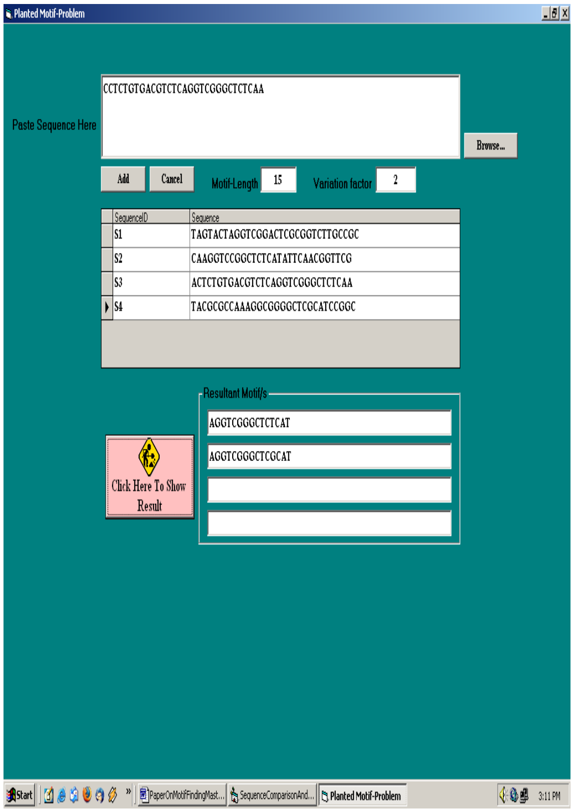
3.3 IMPLEMENTATION

The above described method is implemented as one of the modules in our general purpose computational tool by the name "*Sequence Comparison and Analysis Tool*" for solving various sequence comparison problems encountered in the filed of bioinformatics. It is implemented using Visual Basic-6 package. It has the following functions

a) **SequenceEntry (*QrySeq as string*):** adds the input sequence through interface to he database table 'MotifDB'.
b) **PatternExtractor(*QrySeq as string, WindowSize as integer)* as string**: It extracts all the pattern of size given by window size and stores them in database.
c) **Generate_FATtree(*SeqId* *as* string, *PatternRS as* Recordset)**: generates the FAT tree corresponding to all the patterns of a given sequence.
d) **PatternSearch(*PatternID* as string, *PatternTreeCode as string) as* long**: searches for all the patterns in first sequence in all the remaining set of input sequences.
e) **GenerateMotif *(frequencyDB as* recordset*) as* string**: it generates the required motifs from he frequency table constructed during pattern search phase.

4. Alternative Search Approach

As an alternative approach to search method employed we can extract L-length patterns from all the input sequences in the same manner it is done for first sequence and then store these patterns in relational format. Now we can make use of SELECT…FROM… WHERE pattern LIKE constructs embedded within a procedural code to match most similar patterns corresponding to all the sequences and finally evaluate the frequency at each position as stated in the proposed method to give an approximate motif. The success of this method will depend on the effectiveness of the heuristic employed.



Planted Motif-Problem

Paste Sequence Here: CCTCTGTGACGTCTCAGGTCGGGCTCTCAA

Motif-Length: 15

Variation factor: 2

SequenceID	Sequence
S1	TAGTACTAGGTCGGACTCGCGGTCTTGCCGC
S2	CAAGGTCCGGCTCTCATATTCAACGGTTCG
S3	ACTCTGTGACGTCTCAGGTCGGGCTCTCAA
S4	TACGCGCCAAAGGCGGGGCTCGCATCCGGC

Resultant Motif/s:
AGGTCGGGCTCTCAT
AGGTCGGGCTCGCAT

Figure 3: Interface that captures the input details and shows the output.

5. CONCLUSION AND FUTURE STUDY

Most of the existing methods require some additional information other than the sequences themselves like *SP-STAR, Expectation Maximization methods* or makes some assumption before hand like *Gibbs sampling*. The proposed method makes no prior assumptions or requires additional information about the sequences; it uses a simple algorithm based on dynamic programming approach to determine the motifs from any given number of input sequences of any length. Unlike CONSENSUM method, which is based on greedy algorithm, ours is based on dynamic programming model. As we all know chances of failure in applying greedy algorithms is always quite high in comparison to dynamic programming method.

But we still believe that a more efficient technique can be devised to solve the planted-motif problem using some deterministic algorithm in polynomial time. The use of pipelines in the context of parallel processing can be very handy for providing the solution to the above stated problem. We are already working in this direction and hope to come with a better solution using pipelines.

REFERENCES

[1] H. Leung and F. Chin. Algorithms for Challenging motif problems. *JBCB*, 43—58, 2005.

[2] Stormo, G. DNA binding sites: representation and discovery. Bioinformatics, 16:16{23, 2000.

[3] G.Z. Hertz and G.D. Stormo. Identification of consensus patterns in unaligned dna and protein sequences: a large-deviation statistical basis for penalizing gaps. The Third International Conference on Bioinformatics and Genome Research, 201—216, 1995

[4] Lawrence C E, Altschul S F, Boguski M S, Liu J S, Neuwald A F and Wootton J C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214, 1993.

[5] Cardon, L.R. and Stormo, G.D. (1992) An Expectation-Maximization (EM) Algorithm for Identifying Protein-Binding Sites with Variable Lengths from Unaligned DNA Fragments. J. Mol. Biol. 223:159-170.

[6] Pevzner P A and Sze S H. Combinatorial approaches to finding subtle signals in DNA sequences. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB 2000), 269-278, 2000.

[7] Makino, K. and Uno, T. (2004) 'New algorithms for enumerating all maximal cliques', *Proceedings of Scandinavian Workshop on Algorithm Theory*, pp.260–272.

[8] Rajasekaran, S., Balla, S. and Huang, C.H. (2005) 'Exact algorithm for planted motif challenge problems', *Proceedings of Asia-Pacific Bioinformatics Conference*, pp.249–259

[9] Styczynski, M.P., Jensen, K.L., Rigoutsos, I. and Stephanopoulos, G.N. (2004) 'An extension and novel solution to the (*l*,*d*)-motif challenge problem', *Genome Informatics*, Vol. 15, pp.63–71

[10] Sze S H, Lu S and Chen J. Integrating sample-driven and pattern- driven approaches in motif finding. Lecture Notes in Computer Science/Lecture Notes in Bioinformatics (WABI 2004), 438-449, 2004

Mining Techniques for Integrated Multimedia Repositories: A Review

Naveen Aggarwal[1], **Dr. Nupur Prakash**[2] and **Dr. Sanjeev Sofat**[3]

Abstract - ***The multimedia databases and the need to intuitively handle their content, which meets the user's requirements with the available content based video indexing and retrieval technology, are the main focus of the research in the field of multimedia and computer vision. The researchers mainly focus on the problem of bridging the "semantic gap" between a user's need for meaningful retrieval and the current technology for computational analysis and description of the media content. It takes into account both the high complexity of the real-world implementation and user's need for conceptual video retrieval and browsing. In this paper, the initial work done in this area during last 15 years has been categorized in three generations. The key technologies in each generation are reviewed and characterized based on the standard parameters. It is found that in first and second generation all techniques are semantic less techniques, but in third generation, techniques based on semantics have been evolved. But still most of the techniques are in their infancy and require lots of research for their use in daily applications. In last a solution is proposed for a general purpose multimedia mining application which caters to the needs of the different types of domains.***

1. INTRODUCTION:

The development of various multimedia compression standards in last decade has made the widespread exchange of multimedia information a reality. Due to significant increase in desktop computer performance and a decrease in the cost of storage media, extraordinary growth of multimedia information in private and commercial databases has been seen. Further its ubiquity throughout the World Wide Web, presents new research challenges in computing, data storage, retrieval and multimedia communications. Intuitive handling of this vast multimedia information is the demand of users. Keeping this in mind, multimedia and computer vision researchers are focusing on the development of content based multimedia indexing and retrieval. However, evolution of functional multimedia management system is hindered by the "semantic gap"; a discontinuity between simplicity of content description that can be currently computed automatically and the richness of semantics in user's queries posed for media search and retrieval [1]. The availability of cost effective means for obtaining digital video has led to the easy storage of digital video data, which can be widely distributed over networks or storage media such as CDROM or DVD. Unfortunately, these collections are often not catalogued and are accessible only by the sequential scanning of the sequences. To make the use of large video databases more feasible, user should be able to automatically index, search and retrieve relevant material. Content-Based Video Indexing and Retrieval (CBVIR) has been the focus of the research community during last 15 years. The main idea behind this concept is to access information and interact with large collections of videos for referring and interacting with its content, rather than its form. Although there has been a lot of effort put in this research area, the outcomes have not been very encouraging. The discontinuity between the available content descriptions like color layout or motion activity and the user's need for rich semantics in user queries makes user approval of automated content retrieval systems very difficult. Thus, in order to develop a meaningful CBVIR system one has to involve multidisciplinary knowledge ranging from image and video signal processing to semiotic theories and video production techniques. Signal processing and computer vision methodologies achieved astonishing results in extracting structural and perceptual features from the video data. Algorithms from database system theory and other computer science disciplines enabled efficient, adaptive and intelligent indexing and retrieval of data with various structure and content. Furthermore, fields like computational linguistics and even semiotics have engaged with problems of natural language and even visual media semantics. However, this knowledge is scattered and needs a way to fuse into one system that will enable content-based retrieval of videos in a way natural for users. Multimedia Mining has evolved immensely in last decade. This paper classifies the evolution of Multimedia Mining in three generations. The performance analysis of algorithms of shot detection is analyzed and the paper is concluded with some comments on the future directions.

2. FIRST GENERATION

In the first generation of visual retrieval systems, feature descriptors of video data are expressed manually. Representation of these features provides high level of image ABSTRACTion and model visual content at a conceptual level. These features identify significant entities contained in the image or video (an object, a person, etc.), object parts (eyes in the face, boat in the lake, etc.) or the scene represented and concepts associated to it (a landscape, a storm, etc.). Features can be represented in schemes like relational models and object oriented models and can be queried using query languages like SQL. In Figure 1, main process of first generation is shown. Classification and indexing depends upon how accurate the

[1]University Institute of Engg. & Technology, Panjab University, Chandigarh' H.No 230, Punjab Engineering College Campus, Sector-12, Chandigarh-160012, India. Phone: +91-9814865455
[2]School of Information Technology, Guru Gobind Singh Indraprastha University, Delhi
[3]Computer Sc. & Engg. Deptt., Punjab Engg. College, Chandigarh.
E-Mail: [1]navagg@gmail.com

image features are annotated manually. Cost of annotation is typically very high and the whole process suffers from subjectivity of descriptions, in that the annotator is a different person from the one who issues the query. Search engines like Google and Yahoo uses semantics of web to provide high level descriptions of video data.

2.1 Temporal Video Extraction

In first generation, main concern was to effectively extract the data temporally from video sources. Further video sources may be in compressed or uncompressed domain. All temporal video parsing techniques that exploit information in uncompressed domain lack efficiency. The reason for that is in the nature of the approach. In the feature extraction part the majority of uncompressed analysis techniques must initially decode the video stream and afterwards apply some processing on the vast pixel data, which additionally slows down the processing time. Thus, algorithms that base their analysis on pixel data require substantial processing time. Block-based algorithms [2, 3] and methods based on histogram comparison [4, 5] achieved considerable improvement in both processing requirements and sensitivity to camera and object motion, but far from the efficiency of the compressed domain analysis.

Gargi et al[5] compared the approaches of Arman et al.[6], Patel and Sethi [7], Yeo and Liu [8] and Shen and Delp [9] using different parameters such as: classification performance (recall and precision), full data use, ease of implementation, source effects. Ten MPEG video sequences containing more than 30,000 frames connected with 172 cuts and 38 gradual transitions are used as an evaluation database. It is found that the algorithm of Yeo and Liu and those of Shen and Delp perform best when detecting cuts. Although none of the approach recognizes gradual transitions particularly well, the best performance is achieved by Shen and Delp[9]. The reason for the poor gradual transition detection of the algorithms is their design.

The algorithms are designed by keeping the ideal behavior of transitions in mind. The gradual transitions ideally remain linear in space and time. But the actual frame differences do not follow this ideal pattern smoothly for the entire transition due to the presence of noise. Frame differences are also affected by the degradations due to different sampling and bit rates. Another interesting conclusion is that performance decreases significantly if we don't process all frame types (e.g., like in the first two methods). The algorithm of Yeo and Liu is found to be easiest for implementation as it specifies the parameter values and even some performance analysis is already carried out by the authors. The dependence of the two best performing algorithms on bitrate variations is investigated and shown by Gargi et al [5] that they are robust to bitrate changes except at very low rates. It is found that software encoder implementations also affect the performance. Yeo and Liu compared the effect of different software encoders. Although the techniques that works well for both uncompressed and compressed videos can be considered. But these techniques lack efficiency in either of domain. On the other hand, algorithms that access compressed domain features without additional processing and thus having the similar efficiency, underperformed in the accuracy and robustness criteria.

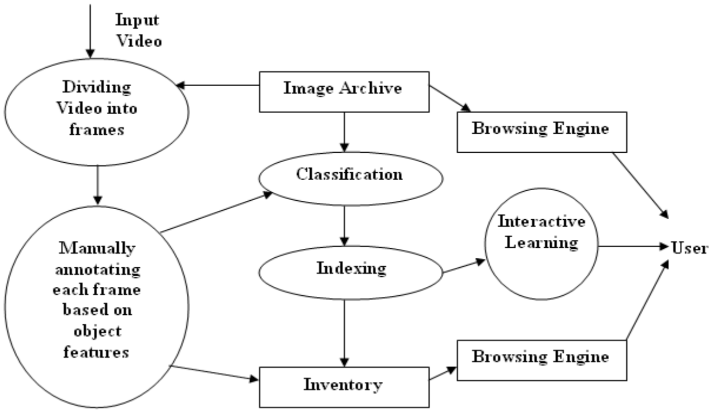


Figure 1: First Generation Video Retrieval Systems

3. SECOND GENERATION

Apart from the descriptors of first generation systems, the second generation systems also describes the perceptual features like color, textures, shape, spatial relationships, etc. These features are numeric descriptors of a video and can be obtained by fully automated objective measurements of the visual contents. So retrieval of content based data can be supported by combination of these features. Most of the techniques used to extract visual primitives from image frames have come from field of computer vision and pattern recognition. Therefore image processing, computer vision and pattern recognition subsystems are integral part of the architecture and operations of the second generation systems. Retrieval is based on similarity models that somehow replicate the way in which humans assess similarity between different objects. Apart from these parameters, Videos can be considered as a source of multi-planar visual information. Each plane communicates different attributes of information. These include the way in which the frames are linked together by using editing effects (cut, fades, dissolves, mattes, etc.), and high level information embedded in the frame sequence (the characters, the story content, the story message). Text embedded in the video frames and the other sensory data like speech and sound can be employed to extract useful data.

Main concern of research on second-generation systems is to extract video structure automatically [10]. In second generation, video indexing is performed by temporally segmenting the videos in unit like shots and scenes. Different image processing and computer vision techniques are used on index frames to generate low level feature descriptors. A metadata database of index frames and their corresponding feature descriptors can be created for later retrieval. When user makes a query, query is transformed into the structurally same low-level feature descriptor and the search engine finds the closest match from a metadata base. The system learns from relevance feedback from users during retrieval process and adapts the feature descriptor in order to achieve more consistent results in terms of perceptual similarity.

Even though some efficient results are reported in literature, there is a problem of bridging the gap between the systems and users. Similarity of perceptual properties is generally of little use in most practical cases of retrieval by content, if not combined with similarity of high-level information.

3.1 Low Level Features

Indexing of images and video contents using low level visual features is touch upon by several content based information retrieval systems. (e.g. WEBSEER [11], QBIC [12], and VisualSeek [13]). Three main task performed by these systems are:

1. Automatic extraction of visual feature descriptors
2. Indexing the extracted descriptors for fast access
3. Querying and matching descriptors for the retrieval of the visual data.

Apart from these features, learning of system is considered as an important aspect. Relevance feedback from users is used to refine the queries and learn through examples that the user may be looking for [14]. More recently, there has been focused effort on automatically producing certain semantic labels that could contribute significantly to retrieving visual data. For example, recent work has focused on portrait vs. landscape detection, indoor vs. outdoor classification, city vs. landscape classification, sunset vs. forest classification [15], [16], and other attempts to answer basic questions of who, what, when, and where about the visual content. Most of the approaches rely on traditional machine learning techniques to produce semantic labels, and some degree of success has been reached for various constrained and sometimes skewed test sets. However, these efforts represent only a small initial step towards achieving the real understanding of the visual content.

3.2 Bridging Gap

Numerous papers [17][18][19][20] have explored the problem of semantic gap and give valuable insights into the current state of the art. Wang et al [17] generate a code book by using the color-texture classification. Different regions of image are segmented using class definitions in the code book. The Entropy of a region contents describes its perceptual importance. Denman et al [18] present the system for creating semantically meaningful summaries of broadcast Snooker footage. Different tools in system are used for parsing the video sequence, identifying relevant camera views, and tracking ball movements. A system for recognizing objects in video sequences is presented by Visser et al [19]. They use the Kalman filter to obtain segmented blobs from the video, classify the blobs using the probability ration test, and apply several different temporal methods, which result in sequential classification methods over the video sequence containing the blob. An automated scene matching algorithm is presented by Schaffalitzky and Zisserman [20]. Main process is base on template matching of given scene in 3-D movies. Ruiz-del-Solar and Navarrete [21] present a system that uses self-organizing maps (SOMs) and user feedback for content-based face retrieval. A ranking algorithm using dynamic clustering for content-based image retrieval is proposed by Park et al [22].

3.3 Shot Detection

Shot Detection technique divides the continuous video, based on recognition of different frame sequences. One shot is characterized when camera remains still i.e. background is still and foreground is changing. When shot changes, background changes significantly. Earlier work in this category is evaluated in [5]. Earlier work is based on recognizing shot boundaries using color histograms, average pixel difference and histogram similarity matrix. Mihai Datcu et al used the motion compensation vector to estimate the motion of different objects [23]. Error in the motion estimation is used to differentiate the key frames from videos. Timo Volkmer described moving query window technique for TREC-12 video techniques [24]. Separate decision stages for abrupt and gradual transitions are applied during a single pass through the video clip. Similarly Masaru Sugano et al performed the shot boundary determination from I picture sequence of MPEG coded video [25]. Both pixel differences and histograms methods are used to overcome problems when either one of them is used. In this same approach is extended to detect shot boundaries in one frame unit. Jordi Mas et al used the color histogram differences and temporal color variations for video shot boundary detection [26]. The technique is able to differentiate abrupt shot boundaries by analysis of color histogram differences and smooth boundaries by temporal color variation. Five different algorithms for shot boundary detection have been analyzed using test data set of different types of videos ranges from news, sports, televisions and movies.

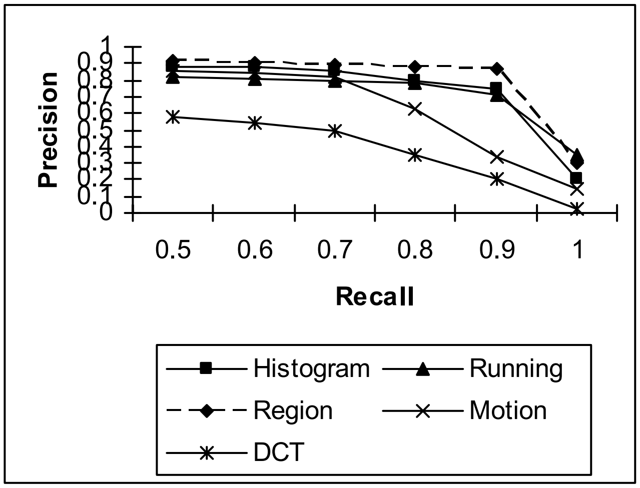


Figure 2: Comparison of Shot Detection Algorithms

As shown in Figure 2, different algorithms have been compared on a common platform from TRECVID 2006 [39] using different set of videos such as BBC News Videos, Sports Video, MTV Videos etc. Two main parameters precision and recall are calculated for each of the algorithm. From graph, it is clear that histogram based and region based algorithms outperforms other in shot detection. Region based algorithm also show some stability in detecting the gradual transitions.

4. THIRD GENERATION

The third generation retrieval systems seek for more high level information from images, audio and video content. Moving videos are not perceived as collection of shots by humans. Spectators even do not realize any editing in video based on the gradual transitions of one shot into another. So our aim is not only to extract the syntactic content (perceptual features) but also the semantic contents. Semantic content involves (1) the rhythm of the sequence (which is induced by the editing), (2) the scenes (which are obtained from shots), (3) the story (including the characters, their roles, actions and their logical relations), and (4) the feelings (which depends on the combination of perceptual facts like colors, objects, music, sounds, etc. and from the meaning of the story). Semantic contents and the perceived feelings of the user are used to support semantic-based retrieval automatically, with no or minimal manual intervention [27]. As human beings are more concerned with the narrative and discourse structure of video contents, so retrieval of video is generally meaningful only if performed at high levels of representation. Image sequence classification must also be based on semantically meaningful categories of information.

4.1 Object Detection

Most challenging problem in visual information retrieval is recognizing and detecting the objects in the moving videos. Several papers present the advances in this area [28][29][30][31]. Matas et al [28] present a method to recognize the objects even when viewing range is wide and illumination is not constant. Their method is robust to occlusion and background clutter in the frames. Sebe and Lew [29] have found that Gaussian noise distribution assumption is invalid. They proposed to use other metrics close to real noise distribution such as Cauchy metric. Further they explained how to create a maximum likelihood metric based on the real noise distributions and found that it consistently outperformed other analytic metrics. A hierarchical shape descriptor for object-based retrieval is proposed by Leung and Chan [30]. Brucale, et al [31] propose a class of topological-geometrical shape descriptors called size functions. Main limitation of this approach is hand drawn shape descriptors, which makes whole idea very much subjective to quality of hand drawn shapes. Based on the idea that the distribution of colors in an image provides a useful clue for image retrieval and object recognition, Berens and Finlayson [32] propose an efficient coding of three dimensional color distributions for image retrieval. Main focus in object recognition area is shifting to make whole process less computational intensive for real time queries.

4.2 Story Segmentation

Story Segmentation is another emerging area in which video is segmented based on different stories. So whereas Shot detection is semantic less technique, story segmentation is semantic based technique. Lekha Chaisorn et al presents a story segmentation technique as shown in Figure 3 using various low level and high level features[33]. It also makes use of various object level features.

4.3 Frameworks

A variety of different frameworks has been proposed to store retrieve and mine the raw videos. Jung Hwan Oh et al proposed such a framework, in which first data is characterized into different segments then each segment is clustered based on low level features and finally data is grouped using motion vectors and multi level hierarchical techniques [34].

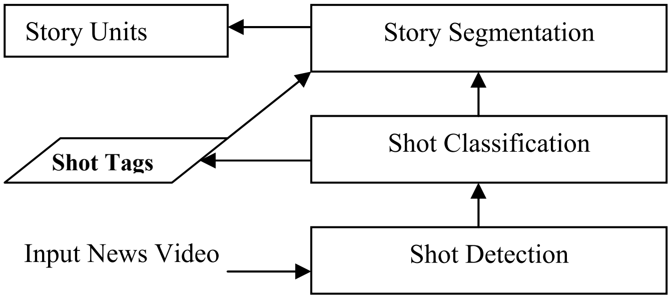


Figure 3: Story Segmentation

It is found that this technique is useful in limited domain and very time consuming. Mihai Datcu et al give an innovative concept for image information mining [23]. It represents the image in four steps. First step is to extract image features using different algorithms. In second step these are grouped together and further data is reduced by parametric modeling of clusters and finally supervised learning is used to represent the knowledge. William Perrizo et al present the significance of P-tree in mining any multimedia data [35]. In this paper, properties of P-trees are presented to take P-tree as Data Mining ready structure. The various implementation issues and scalability of P-tree for large size data are discussed. Xin Huang et al proposed a multimedia data mining framework. It incorporates Multiple Instance Learning into the user relevance feedback in a seamless way. Main aim is to discover the concept patterns of users (especially where the user's most interested region) and how to map the local feature vector of that region to the high-level concept pattern of users [36]. This underlying mapping can be progressively discovered through the feedback and learning procedure. Ankur Teredesai et al proposed the auto-annotation problem as a multi-relational association rule mining problem where the relations exist between image-based features, and textual annotations [37]. Their approach combined low-level image features, such as color, orientation, intensity, etc. and the corresponding text annotations to generate association rules across multiple tables using multi-relational association mining.

5. CONCLUSION

It is clear that Multimedia mining has evolved phenomenally in the last ten years as seen from the progress from first generation to third generation. Whereas, in first and second generation all techniques were semantic less, in third generation, techniques based on semantics have been evolved. But still most of the techniques are in their infancy and require lots of research for their use in daily applications. Multimedia

Mining is being used in different application domain such as Medical diagnosis, News Video Analysis, Sports Video Analysis, Movie Indexing and Satellite image Analysis etc. A general purpose system is still not available which could cater to the need of different domains. Further work need to be done in following areas

1. Multimedia Repositories are supposed to have data in both compressed and uncompressed form. Therefore a representation technique is required which will be able to represent both compressed and uncompressed data in common format which can further be used for mining.
2. Various Multimedia databases provide techniques to organize the data based on its general features. However effective organization should be content based. So a content based organization approach for multimedia data is required.
3. High level features are although based on the semantic of data but they can not be used directly for querying. So an approach is required to relate the semantic of High level feature with the internal representation of data. High level features are usually represented either in First Order Logic or by using Association Rules. Both techniques need to be analyzed in variety of different domains.
4. Further the problem of multi-resolution, scalability and hierarchical video description need to be tackled to gain efficiency in task of video indexing.
5. Optimization of different techniques for temporal extraction of data from compressed and uncompressed domain is required. Already proposed different techniques need to be analyzed and tested with the Benchmark videos proposed by TRECVID.
6. Overall Mining process of Multimedia repositories is bound to be computational and storage expensive. Hence different optimizations can be proposed for limited domains.

Thus a large scale system that merges the semantic text based retrieval approach to multimedia database with content based feature analysis and investigates the significant links between them appears to be the next milestone of research in the field of multimedia management systems [38].

REFERENCES

[1] J. Calic , N. Campbell , A. Calway , M. Mirmehdi , B. T. Thomas, T. Burghardt , S. Hannuna, C. Kong , S. Porter , N. Canagarajah , D. Bull, "Towards Intelligent Content Based Retrieval of Wildlife Videos", Proc. to the 6th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS'2005, EPFL, Montreux, Switzerland, April 2005.

[2] Shu-Ching Chen, Mei-Ling Shyu, Chengcui Zhang, and R.L. Kashyap, "Video Scene Change Detection Method Using Unsupervised Segmentation and Object Tracking," Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), August, 2001, Waseda University, Tokyo, Japan.

[3] Yazdi, M.,Zaccarin, A., Scene break detection and classification using a block-wise difference method,in Proceedings of International Conference on Image Processing, vol 3,Page(s):394 - 397,2001

[4] J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy and N.E. O'Connor, "Temporal Video Segmentation for Real-Time Key Frame Extraction", ICASSP 2002, Orlando, Florida, USA

[5] U.Gargi, R.Kasturi, S.Antani, "Performance Characterization and Comparison of Video Indexing Algorithms" in proceeding of conference on computer vision and pattern recognition (CVPR), 1998

[6] F.Arman, A.Hsu, M-Y Chiu, "Image Processing on Compressed data for large Video databases" in proceedings of First ACM international conference on Multimedia, 1993

[7] ", published in Pattern Recognition N.Patel, I.K.Sethi, "Video Shot detection and Characterization for video databases vol 30, 1997

[8] B.Yeo, B.Liu, "Rapid Scene analysis on compressed video", IEEE Transactions on circuits and systems for video technology, vol 5, no. 6, December, 1995

[9] K.Shen, E.Delp, "A fast algorithm for video parsing using MPEG compressed video sequences", in proceedings of International conference on Image Processing (ICIP '96), 1996

[10] J.Boreczky, L.A.Rowe, "Comparison of video shot boundary detection techniques"in proceedings IS & T/SPIE International Symposium on Electronic Imaging, San Jose 1996

[11] C. Frankel, M. Swain & V. Athitsos, "WebSeer: an image search engine for the world-wide web" Technical Report 94-14, Computer Science Department, University of Chicago, August, 1996

[12] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yonker. Query by image and video content: The QBIC system, in Computer, Vol 28, pages 23–32, 1995.

[13] J. R. Smith and S.-F. Chang. "VisualSEEK: a fully automated content-based image query system." In proceedings of ACM Multimedia '96, November, 1996.

[14] Selim Aksoy, Ozge Cavus, "A Relevance Feedback Technique for Multimodal Retrieval of News Videos," in EUROCON, Belgrade, Serbia & Montenegro, November 21-24, 2005.

[15] M.SZummer, R.Picard, "Indoor-Outdoor Image classification" IEEE Workshop on Content-based access of image and video databases in conjunction with ICCV'98, Bombay, India , 1998

[16] A. Vailaya, A.Jain, H.J.Zhang, "On Image Classification City vs Landscape", IEEE Workshop on Content-Based access of image and video libraries", Santa Barbara, CA, June, 1998

[17] Wang,W., Song, Y., Zhang.A "Semantics-based image retrieval by region saliency", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer 2002.

[18] Denman, H., Rea, N., Kokaram, A "Content based analysis for video from snooker broadcasts" in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[19] Visser, R., Sebe, N., Bakker, E, "Object recognition for video retrieval", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[20] Schaffalitzky, F., Zisserman.A "Automated scene matching in movies. In: International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[21] Ruiz-del-Solar, J., Navarrete, P, "FACERET: An interactive face retrieval system based on self-organizing maps" in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[22] Park, G., Baek, Y., Lee, H.K, "A ranking algorithm using dynamic clustering for content-based image retrieval", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[23] Mihai Datcu and Klaus Seidel, "An Innovative Concept For Image Information Mining", in proceeding of International workshop on Multimedia Data Mining with ACM SIGKDD, 2002

[24] Timo Volkmer, S.M.M.Tahaghoghi, Hugh E. Williams, James A. Thom, "The Moving Query Window for Shot Boundary Detection at TREC-12", Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, October 2003.

[25] Masaru Sugano, Keiichiro Hoashi, Kazunori Matsumoto, Fumiaki Sugaya, Yasuyuki Nakajima, "Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2003" ,

[26] Jordi Mas, Gabriel Fernandez, "Video Shot Boundary Detection based on Color Histogram", Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, October 2003.

[27] Osmar R. Zaiane, Jiawei Han, Ze-Nian Li, Jean Hou, "Mining Multimedia Data", MDM/KDD2001, Second International Workshop on Multimedia Data Mining in conjunction with Seventh ACM SIGKDD International Conference on Knowledge Discovery & Data Mining , San Francisco, CA, USA, 2001

[28] S. Obdrˇz´alek, Matas, J, "Local affine frames for image retrieval", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[29] Sebe, N., Lew, M, "Robust shape matching", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[30] Leung, M.W., Chan, K.L, "Object-based image retrieval using hierarchical shape descriptor. ", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[31] Brucale, A., d'Amico, M., Ferri, M., Gualandri, M., Lovato, A., "Size functions for image retrieval: A demonstrator on randomly generated curves", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer 2002

[32] Berens, J., Finlayson, G, "An efficient coding of three dimensional colour distributions for image retrieval", in International Conference on Image and Video Retrieval, Lecture Notes in Computer Science, vol. 2383, Springer, 2002

[33] Lekha Chaisorn, Tat-Seng Chua, Chun-Keat Koh, Yunlong Zhao, Huaxin Xu, Huamin Feng, Qi Tian, " A Two-Level Multi-Modal Approach for Story Segmentation of Large News Video Corpus", Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, October 2003.

[34] JungHwan Oh and Babitha Bandi, "Multimedia Data Mining Framework For Raw Video Sequences", in proceeding of International workshop on Multimedia Data Mining with ACM SIGKDD, 2002

[35] William Perrizo, William Jockheck, Amal Perera, Dongmei Ren, Weihua Wu, Yi Zhang, "Multimedia Data Mining Using P-Trees", in proceeding of International workshop on Multimedia Data Mining with ACM SIGKDD, 2002

[36] Xin Huang, Shu-Ching Chen, Mei-Ling Shyu and Chengcui Zhang, "User Concept Pattern Discovery Using Relevance Feedback And Multiple Instance Learning For Content-Based Image Retrieval", in proceeding of International workshop on Multimedia Data Mining with ACM SIGKDD, 2002

[37] Ankur Teredesai, Juveria Kanodia, Muhammad Ahmad, Roger Gaborski, "CoMMA: A Framework for Integrated Multimedia Mining using Multi-relational Associations", in proceeding of International workshop on Multimedia Data Mining with ACM SIGKDD, 2004

[38] Osmar R. Zaïane, Simeon J. Simoff, Chabane Djeraba, "Mining Multimedia and Complex Data", KDD Workshop MDM/KDD 2002, PAKDD Workshop KDMCD 2002, Revised Papers, ISBN 3-540-20305-2, Springer 2003

[39] Evaluation tools, campaigns and TRECVID Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, http://www-nlpir.nist.gov/projects/t01v/trecvid.tools/shot.boundary.evaluation/sbeval/, 2006

Antecedents of Success in Business Process Outsourcing: An Empirical Study of the Indian BPO Sector

Aparna Daityari[1], **Prof. A. K. Saini**[2] and **Romit Gupta**[3]

Abstract - Global sourcing of technology related services is estimated to have grown to reach USD 70-76 billion in 2007 of which IT-BPO services, growing at an above-sector-average rate of nearly 8 per cent, remain the largest category, accounting for an increasing share of the worldwide technology sector revenue aggregate (Nasscom, 2008). This rising ubiquity of IT enabled outsourcing especially business process outsourcing (or BPO) has led to heightened interest in both the practitioner and academic communities on the reasons and practices leading to successful outcomes. Both communities have increasingly referred to business process outsourcing as a form of strategic partnership indicating the predominance of relational governance mechanisms. Reality however is far from that. We establish the importance on control over partnership quality exemplified in the form of a structured contract and its associated monitoring mechanisms through an empirical study of 124 business process outsourcing relationships in some of India's most established BPO vendors.

Index Terms - Business process outsourcing, control, management, partnership quality, outsourcing success

1. INTRODUCTION

The three trends that seem certain to dominate the world, for some time to come, are globalization, technological advances and deregulation. They combine to make geographical dispersion an area of low concern in the planning of business strategy; as enterprises increasingly look for leveraging the cost or differentiation advantages available across the globe - forging partnerships to create a value chain with the aim of accomplishing the most with the least. It is in this scenario that business process outsourcing ('bpo') has emerged as a key component of management strategy as a global supply-chain of information and expertise that stretches from Mumbai to Manhattan is etched.

Today, as outsourcing is enmeshed into most organizational strategic plans; relationship management issues have become central. As Gottfredson et al. state 'the question is no longer whether to outsource a capability or activity but rather how to source every single activity in the value chain' (Gottfredson et al., 2005). The management of the relationship includes all conscious activities of the parties to impact the relationship during its life in their desired way, and is constructed around two main elements: (a) the formal contract and the associated governing mechanisms viz. *control mechanisms* and (b) the 'psychological contract' (Sabherwal, 1999) that is based on the parties' mutual beliefs and attitudes and is reflected as *partnership quality*. There is a natural link between how an outsourcing arrangement is structured and managed, and the subsequent outcomes (Dibbern et al., 2004).

In this paper, we have looked at this issue of governance; specifically we study empirically the dynamics of interaction of degree of control and partnership quality of a business process outsourcing relationship and it's impact on outsourcing success. Our study site is the business process outsourcing sector of India as this country is considered the Write the body of the paper here. undisputed leader' (A T Kearney, 2006) of this sector. The overall Indian IT-BPO revenue aggregate is expected to grow by over 33 per cent and reach USD 64 billion by the end of the current fiscal year (Nasscom, 2008).

2. REVIEW OF LITERATURE.

Business process outsourcing has often been treated as an extension of the concept of IT/IS outsourcing to IT intensive business processes (Hyder et al, 2002) or as a subset of IT/IS outsourcing (Sovie & Hansen, 2001; Michell & Fitzgerald, 1997) leading to "a virtual absence of academic publications on the topic" (Rouse and Corbitt, 2004). Other researchers such as Gewald et al (2006) and Whitaker et al. (2006) have also noted this lacuna. Hence for this paper, we have sourced information from the extensive body of work on IT/IS outsourcing research of the past two decades. As Dibbern et al. (2004) note "research on ………. business process outsourcing would benefit from 'standing on the shoulders' of what has already been accomplished in the field of IS outsourcing".

2.1 Business Process Outsourcing

The transaction cost approach to the theory of the firm hypothesizes that firms are organizational innovations born out of the costs involved in market transacting in order to reduce those costs. Coase (1937) has argued that, were the firm and the market alternatives for organizing the same set of transactions; a firm will substitute market transactions as long as management costs are less than transaction costs. Thanks to the convergence in corporate computing platforms and rapid advances made in communications technology it has become easy and inexpensive to seamlessly link together geographically dispersed information systems thus making market transactions for executing several activities previously done within the firm boundaries possible and preferable. This concept of remotely executing tasks was the genesis of

[1]Associate Prof.- BLS Institute of Management, Ghaziabad (India), Research Scholar-GGS Indraprasth, University, Delhi (India), [2]Professor of Management- GGS Indraprastha University Delhi (India), Mobile:00919811165001
[3]Country Head-Wipro BPO Philippines Inc.
E-Mail : [1]aparna_daityari@rediffmail.com,
[2]aksaini@rediffmail.com, [3]gupta.romit@gmail.com

business process outsourcing defined as “the delegation of one or more IT-intensive business processes to an external provider that, in turn, owns, administrates and manages the selected process/processes, based upon defined and measurable performance metrics” (Gartner 2004).

2.2. Business Process Outsourcing In The Indian Context

In 1994 American Express Travel Related Services combined, standardized, and re-engineered its more than 46 transaction processing sites sprawled across North America, Latin America, EMEA (Europe, Middle East, and Africa), and APA (Asia Pacific and Australia) into 3 sites. Besides Phoenix, Arizona and Brighton, United Kingdom, Gurgaon was chosen as the location for the third site dedicated to handling Japan and Asia Pacific and Australia (Kalakota and Robinson, 2004). In the mid 1990s GECIS also opened its captive customer services centre in Gurgaon. The first pure play BPO provider was Spectramind eServices Pvt. Ltd. set up by Mr. Raman Roy in March 2000 and subsequently bought over and renamed Wipro BPO in 2005. Thus, though the sector is more than two decades old, academic research on the Indian BPO sector is dominated by articles analyzing the country’s comparative advantages, especially with respect to China (see for instance (Popkin and Iyengar, 2007) and case studies of successful Indian vendors (Holloway et al, 2006 on EXL; Wendell and Arippol, 2007 on Daksh; Yadav et al, 2006 on Technovate amongst a host of others).

2.3. Control

Control can be defined as the organization's attempt to increase the probability that people will behave in ways that lead to the attainment of organizational goals (Flamholtz et al., 1985) and thus includes a range of mechanisms to monitor and execute operations. These control practices help implement control modes, which may broadly be divided into formal controls, i.e., modes that rely on practices that influence the controllee’s behavior through performance evaluation and rewards, and informal controls, i.e., modes that utilize social or people strategies to reduce goal differences between controller and controllee (Eisenhardt 1985; Kirsch 1996, 1997). Tannenbaum (1968) further proposed that two different types of control “can operate concurrently and that their effect is additive”. Thus Tannenbaum (1968) presented the approach that control types are not mutually exclusive, that their control ‘levels’ are quantifiable and that they can be summed up to one aggregate level.

2.4. Partnership Quality

Lee and Kim (1999) tested the impact of different attributes of partnership quality on satisfaction with the vendor’s IS services. Partnership quality was defined as the degree of trust, business understanding, benefit and risk sharing, commitment, and conflict. Of these factors – with the exception of conflict – partnership quality was found to significantly influenc outsourcing success.

2.5. Outsourcing Success

Grover et al. (1996) proposed that organizations expect to gain some degree of advantage, typically in the form of economic, technological or strategic benefits in an outsourcing relationship. Saunders et al.

(1997) used these three dimensions of benefits (economic, technological and strategic) and added overall satisfaction to determine the outsourcing success for their study. Lee and Kim (1999) explored outsourcing success on two dimensions - business perspective and user perspective. The business perspective focused on the sets of economic, technological and strategic benefits that the organization could achieve through outsourcing whereas the user dimension focused exclusively on user satisfaction with the outsourcing arrangement.

3. RESEARCH METHODOLOGY.

This section presents the methodological elements, the research design and the operationalization of the variables in this study.

3.1. Research Design

Both control (Sabherwal 1999, Marcolin & McLellan 1998, Rustagi 2004) and partnership quality (Grover et al. 1996, Lee & Kim 1999) have been established as antecedents of outsourcing success by earlier researches. As Clark et al. state “….the truly critical success factors associated with successful outsourcing are those associated with vendor governance” (Clark et al., 1998).

The relationship between control and partnership quality has however seen conflicting views in previous research. Some researchers claim that control mechanisms and partnership quality especially it’s component of trust can be pursued simultaneously, or are complementary (Zaheer and Venkatraman, 1995; Das and Teng, 1998), while others argue that control mechanisms are detrimental to trust (see, for example, Lorange and Roos, 1992). In this study it is proposed that the interaction between degree of control and partnership quality is an antecedent of business process outsourcing success. This view is supported by Marcolin and McLellan (1998) whose research indicated that relationships achieved greater satisfaction through more control and certainty in their relationships, and were better in avoiding conflict, achieving cost reductions, and developing trust.

As presented in Figure 1, a two-by-two matrix is proposed to depict the impact of interaction between degree of control and partnership quality on outsourcing success. A high degree of control along with high partnership quality is proposed to facilitate a high degree of success. A high degree of either control or partnership quality along with low degrees of the other should be associated with moderate degrees of success. On the other hand, the combination of both low control and partnership quality should be associated with a lower degree of success in outsourcing.

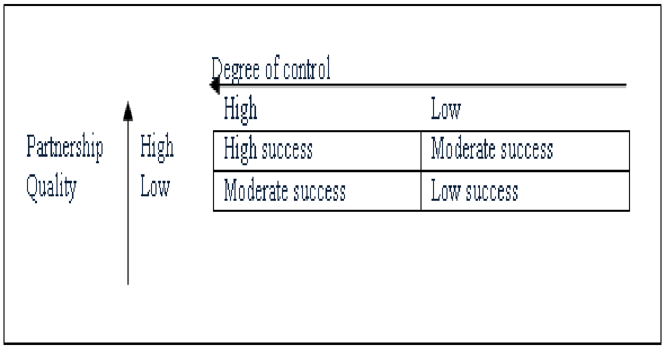


Figure 1: Control, Partnership quality and outsourcing success: The Relationship

Mathematically, the relationship among degree of control, partnership quality and outsourcing success can be depicted as:

S=f(C, PQ, C*PQ)

Where S= business process outsourcing success

C= control

PQ= partnership quality and

C*PQ= interaction between control and partnership quality

Thus, this study presents an argument that the interaction between the degree of control and partnership quality will have a positive relationship with outsourcing success, hypothesized as:

> ***The degree of control and partnership quality would have a positive interactive association with outsourcing success.***

3.2. Research Constraints

A major constraint faced by the researcher was the stringency of data security measures. Possibility of misuse of sensitive and proprietary data coupled with lack of confidence in the Indian regulatory framework for protection of intellectual property are major client concerns, prompting vendors to put in place elaborate security measures which prohibit employees from sharing information without formal permission, restricted access to internet, cell phones, email, instant messaging, removable storage units such as pen drives, laptops etc. Further, fears of backlash and stakeholder concerns in the home country have made clients extremely wary of revealing their outsourcing arrangements. All relationships studied by this researcher had 'non-disclosure agreements' as part of the contract prohibiting the vendor from disclosing any information which would publicly identify the client. As a precondition for allowing access to personnel and information, this researcher was asked to refer any material collected, for review by the vendor's management prior to inclusion in this study. Further, any information identifying either the vendor or the client organization was strictly prohibited.

3.3. Data Collection

As a master list, the "List of 50 best-managed global outsourcing vendors, 2006" from The Black Book of Outsourcing survey, conducted by Brown-Wilson Group, a Clearwater, Fla., based consultancy, was taken. This survey ranked outsourcing vendors according to responses by executives and others involved in outsourcing decision-making about their experience and satisfaction with current suppliers and appeared in 'Sourcing Magazine' July 2006. The survey has wide acceptance and credibility in the global business process outsourcing sector (BusinessWire, June 2007). Of this list permission to interview managerial employees from the tactical and strategic levels from twenty was obtained. The vendor organization management were then requested to solicit the support of their counterpart client teams for this study. The resultant sample had 13 vendor organizations with access to senior managers from both client and vendor sides. Since this process had excluded the captive business process outsourcing centers, access was sought and gained to study 2 captive centers as well. This was based on personal and professional acquaintance, hence a convenience sample.

Thus a final sample of fifteen business process outsourcing vendor organizations was taken. At our request, these organizations identified one or more vendor-client relationships among their ongoing lines of business. The client executives and vendor executives who agreed to participate in the study were sent an email detailing information about the purpose of the study, the level of participation required and potential benefits. It also stressed the anonymity and confidentiality of the respondents. This was followed by interviews over the period August 06- July 07. In sum, we surveyed 124 relationships between 103 service receivers and 15 service providers through 228 interviews.

3.3. Measure Development

Two survey instruments (included in appendices) were developed for this study. The first one was used to collect information from the client executive and the other was used with vendor executives for control practices and partnership quality. The measures were administered through interviews which were semi structured and detailed. Interviews were held individually for one to three hours per session.

Prior to measure development, we conducted a series of personal interviews with seven BPO professionals to confirm the external validity of the developed research framework. The interviews confirmed that our proposed research model was suitable for studying real world outsourcing phenomena. We then developed a five-point Likert-style questionnaire based on the literature and the comments gathered from the interviews for control practices.

3.4. Operationalization Of Variables

1. Degree of control

   Items for the construct of degree of control have been developed by creating a list of control mechanisms, which are then individually operationalized. Kirsch (1997) and the findings of an earlier work by the authors of this research work (Daityari et al., 2007) are the sources for this list.
2. Outsourcing success

This construct measures the extent to which the client organization achieved their outsourcing objectives as assessed through economic benefits, technological benefits and strategic benefits. Items for this construct have been adapted from Grover et al. (1996); Lee and Kim (1999) and Rustagi (2004).

3. Partnership quality

   This construct measures the factors that make up partnership quality viz. trust, business understanding, benefit/risk share and commitment. This construct was measured with an adapted version of the instrument used by Lee & Kim (1999) which employs mostly perceptual measures for the constituent variables.

4. DATA ANALYSIS

4.1. Sample Characteristics

The target population of this study was the client vendor teams belonging to business process outsourcing relationships. The initial list adopted for this study was the fifty top BPO vendors in India from which access to representatives from both client and vendor sides from 13 were obtained (2 were added later from captive firms). Hence the response rate of this study may be taken as 26%.

Among the fifteen vendors who participated in our study, six were ‘IT Outsourcers’, one was a ‘business process specialist’, another two were ‘pure play BPO providers’. ‘Captives’ were represented by two respondents while there were three ‘former captives’ and one ‘BPO consultant’. In terms of vendor size, with respect to employee headcount, the smallest was 3300 and the biggest 22000. Revenues earned from BPO alone for the vendors studied ranged from $54mn to $613mn excluding the captive units.

Majority of the clients (38%) were from the financial services sector, followed by telecom (12%) and IT hardware and software (11%). Contact and front office services (16) comprised the largest chunk of the outsourced processes, followed by financial and accounting (14) and knowledge services (12).

4.2. Reliability And Validity Of The Measurement Instrument

The content validity of the instruments was established through the adoption of the constructs that have been validated by other researchers and a pretest with outsourcing professionals. We calculated the internal consistency (Cronbach's alpha) in order to assess the reliability of the measurement instrument. For convergent validity, we evaluated the item-to-total correlation that is the correlation of each item to the sum of the remaining items. Discriminant validity was checked by means of a factor analysis. Results of these statistical tests are provided here.

Construct	Coding	Item-to-total correlation	Factor loading
Formal control (α= .870)	FC		
	FC1	.844	0.98
	FC2	.665	0.75
	FC3	.942	0.94
	FC4	.858	0.982
	FC5	.040	0.948
	FC6	.921	0.952
Informal control (α= .879)	IC		
	IC1	.780	0.894
	IC2	.908	0.933
	IC3	.781	0.871
	IC4	.616	0.79
Outsourcing success (α= .918)	OS		
	OS1	.749	0.834
	OS2	.841	0.907
	OS3	.892	0.938
	OS4	.824	0.892
	OS5	.658	0.773
Trust (α = .881)	PQT		
	PQT1	.795	.903
	PQT2	.704	.857
	PQT3	.881	.951
Business understanding (α = .991)	PQBU		
	PQBU1	.756	.885
	PQBU2	.864	.941
	PQBU3	.857	.938
Benefit / Risk share (α = .795)	PQBR		
	PQBR1	.712	.925
Commitment (α = .708)	PQBR2	.712	.925
	PQC		
	PQC1	.570	.839
	PQC2	.378	.658
	PQC3	.668	.883

Table 1: Reliability and Validity statistics

4.3. Analysis of Empirical Data

This section presents the statistical tools used to test the hypothesis proposed in the study and presents the results thereof. Our hypothesis stated that the degree of control and partnership quality would have a positive interactive association with outsourcing success.

To test this, first, a hierarchical regression analysis was performed. In the first step of hierarchical regression analysis, control and partnership quality were used as the independent variables with outsourcing success as the dependant variable. Here control was found to have a significant positive relationship with outsourcing success (β=.683, *p*= .000) while that of partnership quality was not significant (*p= .954).*

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.683[a]	.466	.457	2.31927

a. Predictors: (Constant), PQ, CNTRL

ANOVA[b]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	567.688	2	283.844	52.769	.000[a]
	Residual	650.860	121	5.379		
	Total	1218.548	123			

a. Predictors: (Constant), PQ, CNTRL

b. Dependent Variable: OS

COEFFICIENTS[A]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-24.555	3.942		-6.229	.000
	CNTRL	1.168	.114	.683	10.273	.000
	PQ	-.002	.035	-.004	-.058	.954

a. Dependent Variable: OS

Then in the next step of this analysis, the interaction between control and partnership quality was introduced as another independent variable in addition to the two variables entered earlier. The interaction between control and partnership quality also did not indicate any significant relationship with outsourcing success.

MODEL SUMMARY

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.683[a]	.466	.457	2.31927

a. Predictors: (Constant), cntrl_pq, CNTRL

ANOVA[b]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	567.688	2	283.844	52.769	.000[a]
	Residual	650.860	121	5.379		
	Total	1218.548	123			

a. Predictors: (Constant), cntrl_pq, CNTRL

b. Dependent Variable: OS

COEFFICIENTS[A]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-24.555	3.942		-6.229	.000
	CNTRL	1.170	.119	.684	9.809	.000
	cntrl_pq	-.002	.035	-.004	-.058	.954

a. Dependent Variable: OS

EXCLUDED VARIABLES[B]

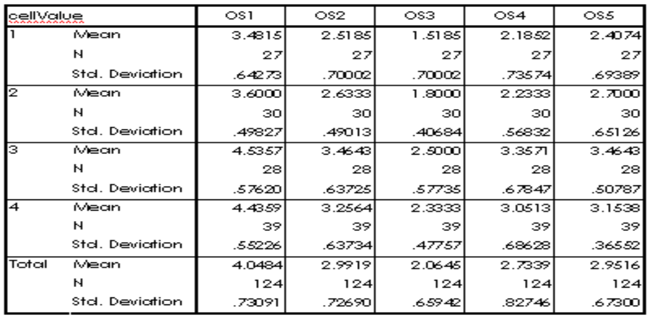
Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	PQ	.[a]	.	.	.	.000

a. Predictors in the Model: (Constant), cntrl_pq, CNTRL

b. Dependent Variable: OS

The next technique used to examine the influence of interaction between the degree of control and partnership quality on outsourcing success was the Median Split Analysis. Using the median values for degree of total control and partnership quality, the 124 BPO relationships were split into four subgroups (cell 1= low/low; cell 2= high PQ/ low Cntrl; cell3= low PQ/ high Cntrl; cell 4= high/high). The average values of outsourcing success variables (OS 1 to OS 5) in the appropriate cell were then examined.

As presented in the table, the top two rows have lower values for outsourcing success than the bottom two rows indicating that level of outsourcing success was significantly associated with the degree of control in the relationship (high control associated with high success). On the other hand, the almost identical nature of the top two rows indicates that level of partnership quality has no significant association with outsourcing success.



cellValue		OS1	OS2	OS3	OS4	OS5
1	Mean	3.4815	2.5185	1.5185	2.1852	2.4074
	N	27	27	27	27	27
	Std. Deviation	.64273	.70002	.70002	.73574	.69389
2	Mean	3.6000	2.6333	1.8000	2.2333	2.7000
	N	30	30	30	30	30
	Std. Deviation	.49827	.49013	.40684	.56832	.65126
3	Mean	4.5357	3.4643	2.5000	3.3571	3.4643
	N	28	28	28	28	28
	Std. Deviation	.57620	.63725	.57735	.67847	.50787
4	Mean	4.4359	3.2564	2.3333	3.0513	3.1538
	N	39	39	39	39	39
	Std. Deviation	.55226	.63734	.47757	.68628	.36552
Total	Mean	4.0484	2.9919	2.0645	2.7339	2.9516
	N	124	124	124	124	124
	Std. Deviation	.73091	.72690	.65942	.82746	.67300

Thus, Median Split Analysis supports the results of Hierarchical Regression Analysis in regard to the significant relationship between the degree of control and outsourcing success.

5. DISCUSSION OF FINDINGS

The hypothesis stated that the degree of control and partnership quality would have a positive interactive association with outsourcing success. This was tested using Hierarchical Regression Analysis and Median Split Analysis. Results of both indicated that while degree of control has a significant positive relationship with outsourcing success neither partnership quality nor its interaction with degree of control enjoyed any relationship of significance with outsourcing success.

Control as a significant predictor of success finds empirical support from research work done on other inter-organizational settings. Henderson and Lee's (1992) study on control behaviors that can affect the performance of an I/S design teams supported the proposition that increases in the total level of control behavior is positively correlated with performance.

Results on partnership quality are however in contradiction with what is most often suggested by earlier research. Lee and Kim's (1999) study stated partnership quality was a critical success factor of IS outsourcing. Further, primarily based on social exchange theory, many researchers argue that a partnership is the ideal type of relationship, in that it will result in a higher level of success (Dibbern 2004).

Support for this result however comes from Lacity and Hirschheim (1993) who revealed that the trade literature presented an inaccurate view of outsourcing arrangements. The outsourcing relationship is often portrayed as a strategic partnership or alliance. This is at odds with the actual contractual relationship, which usually does not contain provisions for sharing risks and rewards associated with outsourcing. Lacity and Hirschheim go on to state that viewing a relationship as a partnership can be dangerous because it may lead to a loose or incomplete contract, in part because the client thinks of the outsourcing vendor as a partner when in fact it is not. Consequently, the possibility for "opportunistic behavior" by the vendor exists.

Even in the Indian context explored by this study, while outsourcing clients and vendors interviewed by us, referred to the outsourcing relationship variously as 'strategic partnership', 'thought partnership' 'enterprise extensions' and in fact most had pricing models based on 'gain share' agreements which reflected the benefit –risk sharing component of partnership quality; the governance of the relationship was structured around formal control mechanisms which were explicitly guided by the contract or 'statement of work'. This reality was well brought out by the Vice President (Operations) of the largest (in terms of revenue) business process outsourcing vendor of our study who referred to the 'Statement of Work' (formal contract) as the "bible" guiding the relationship.

For an outsourcing client, the predominance of degree of control over partnership quality as an indicator of outsourcing success indicates the need to pay adequate attention to establishing a rigorous control structure at the very onset of the relationship. It is also equally important to build in the flexibility to accommodate experiential learning at later stages. The contract or 'Statement of Work' (SOW) is of paramount importance as it is the single most definitive guide to the governance of the relationship. Thus, it is important to be precise in specifying process, performance and people related requirements, install robust post-contract monitoring mechanisms, and to build in a pricing structure which involves the vendor in sharing the benefits and risks of the relationship.

For the vendor, this study underscores the point that in-spite of all the public pronouncements on 'strategic alliances' and 'close partnerships' the ground reality of the Indian BPO sector is that it is largely transactional in nature with the relationships being that of typical buyer-supplier nature. This is emphasized by the fact that this study is focused on the dominant players of this sector. Earlier research has suggested vendor concerns such as inflexibility and traditional measurement approaches of clients as a barrier to service innovation (Quinn, 1999). This indicates the need for vendors to encourage and educate clients to move towards a model of increasing partnership enabled governance as a remedy to their concerns.

CONCLUSION

The purpose of this study was an examination of the antecedents of success of business process outsourcing relationships. A survey methodology was utilized involving matched pair samples of client and vendor executives in business process outsourcing relationships. Overall, degree of control was established as a significant predictor of outsourcing success while partnership quality was not found relevant to the success of the relationship.

This paper makes some significant contributions of interest to managers who enter outsourcing relationships, and to researchers who endeavor to understand the nature of control systems associated with new organizational forms. The first contribution is support for the proposition that a control theories perspective may be helpful in explaining the relative success of business process outsourcing relationship. Further, in existing outsourcing research, the independent variables selected to explain outsourcing success, display a strong focus on the "soft" elements of an outsourcing relationship – often named as partnership attributes (Dibbern et al., 2004). This paper attempts to rectify that imbalance by adopting a comprehensive approach inclusive of both formal governance mechanisms alongside relational mechanisms.

REFERENCES:

[1] A T Kearney's Global Services Location Index 2006

[2] Business Wire : 'Adventity Ranked in Top 25 of Best Managed Global Outsourcing Companies According to Black Book of Outsourcing' June 18, 2007

[3] Clark, T. D., Jr., Zmud, R. W. and McCray, G. E. (1995) "The Outsourcing of Information Services: Transforming the Nature of Business in the Information Industry," Journal of Information Technology, Vol. 10, pp. 221-237.

[4] Coase, R. H. (1937). "The Nature of the Firm," Economica, Vol. 4, No. November, pp. 386-405.

[5] Daityari, A., Saini, A.K. and Gupta, R; 'Portfolios of control in business process outsourcing' 2007, Proceedings of the National Conference on Management, GGSIP University Delhi

[6] Das, T.K. and Teng, B-S. (1998), "Between trust and control: developing confidence in partner cooperation in alliances", Academy of Management Review, Vol. 23, pp. 491-512Dataquest Gartner 2004 available at www.gartner.com

[7] Dibbern, J., Goles,T., Hirschheim,R., Jayatilaka, B., Information Systems Outsourcing: A Survey and Analysis of the Literature; The DATA BASE for Advances in Information Systems - Fall 2004 (Vol. 35, No. 4)

[8] Flamholtz, E., Das,T. and Tsui, A 'Toward an Integrative Framework of Organizational Control', Accounting Organizations and Society, 10-1, 1985, 35-50

[9] Gartner Dataquest 2004, available at *www.gartner.com*

[10] Gewald, H., K. Wüllenweber and T. Weitzel (2006). "The Influence of Perceived Risks on BankingManagers' Intention to Outsource Business Processes - A Study of the German Banking and finance Industry." Journal of Electronic Commerce Research 7(2): 78-96.

[11] Gottfredson, M., Puryear, R. and Phillips, S 'Strategic Sourcing' Harvard Business Review February 2005

[12] Grover, V., Cheon, M.J. and Teng, J.T. 'The effect of service quality and partnership on the outsourcing of information system functions'; Journal of Management Information Systems, 12(4), 1996.

[13] Henderson, J.C. and Lee, S "Managing I/S Design Teams: A Control Theories Perspective" Management Science, 38(6), June 1992

[14] Holloway, C; Morgridge,J.P. and Spitzer, J "ExlService: Business Process Outsourcing in India"; Harvard Business Review, Sep 2006

[15] Hyder, Kumar, Mahendra, Seigel, Heston, Gupta, Mahaboob and Subramanium. 2002. 'eSourcing Capability Model for IT enabled service providers' School of Computer Science; Carnegie Mellon University

[16] Kalakota, R. and Robinson, M. 'Emerging Business Models in Offshore Outsourcing' http://www.sterlinghoffman.com/newsletter/articles/article107.html. May 2004.

[17] Kirsch, L. J. 1996. The management of complex tasks in organizations: Controlling the systems development process. Organ. Sci. 7(1) 1–21. .

[18] Kirsch, L. J. 1997. Portfolios of control modes and IS project management. Inform. Systems Res. 8(3) 215–239.

[19] Lacity, M. C. and Hirschheim, R. A. (1993). "The Information Systems Outsourcing Bandwagon," *Sloan Management Review*, Vol. 35, No. 1, pp. 73-86

[20] Lee, J.-N. and Kim, Y.-G. (1999). "Effect of Partnership Quality on IS Outsourcing Success: Conceptual Framework and Empirical Validation," *Journal of Management Information Systems*, Vol. 15, No. 4, pp. 29-61

[21] Lorange, P. and Roos, J "Strategic alliances: formation, implementation, and evolution" Blackwell Business Publication, 1992

[22] Marcolin, B. L. and McLellan, K. L. (1998). "Effective IT Outsourcing Arrangements," *Proceedings of the 31st Annual Hawaii International Conference on System Sciences*, pp. 654-665

[23] Michell, V. and Fitzgerald, G. (1997). "The IT Outsourcing Market-Place: Vendors and their Selection," Journal of Information Technology, Vol. 12, pp. 223-237.

[24] Nasscom Strategic Review 2008

[25] Popkin, J. M. and Iyengar, P "IT and the East: How China and India Are Altering the Future of Technology and Innovation" Harvard Press, 2007

[26] Quinn, J. 1999. 'Strategic Outsourcing: Leveraging Knowledge Capabilities', Sloan Management Review, 40(4):9-24

[27] Rouse, Anne C. & Corbitt, Brian J. (2004). Business Process Outsourcing: Promises Promises. ABIE Source (Annual). pp 86-87

[28] Rustagi, S. (2004) . Antecedents of success in IS outsourcing: A control theory perspective. Unpublished PhD Thesis. University of Pittsburgh

[29] Sabherwal, R. (1999). "The Role of Trust in Outsourced IS Development Projects," *Communications of the ACM*, Vol. 42, No. 2

[30] Saunders, C., Gebelt, M. and Hu, Q. (1997). "Achieving Success in Information Systems Outsourcing," *California Management Review*, Vol. 39, No. 2, pp. 63-79

[31] Sovie, D., and Hanson, J. (2001): The xSP Revolution, Round 2: Separating Winners From Losers, Mercer Management Consulting, http://www.mercermc.com/Perspectives/WhitePapers/Commentaries/Comm01ASP.pdf

[32] Tannenbaum, A,, (ED,), Control in Organizations, McGraw-Hill, New York, 1968.

[33] Wendell, P.C. and Arippol, P. "Daksh (A): 1999 Business Plan" Harvard Business Review, Feb 2007

[34] Whitaker, J., Bardhan, I.R., Mithas, S., Antecedents of Business Process Outsourcing in Manufacturing Plants; Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 08, 2006, 168.1

[35] Yadav, V.; Bharadwaj, S. and Saxena, K.B.C "Tecnovate: Challenges of Business Process Outsourcing " Harvard Business Review Dec 2006

[36] Zaheer, A and Venkatraman, N "Relational governance as an interorganizational strategy: An empirical test of the role of trust in economic exchange" Strategic Management Journal 16(5), 1995

APPENDICES: SURVEY INSTRUMENTS

APPENDIX A: VENDOR QUESTIONNAIRE

In this research study, the responses from the questionnaires will be aggregated at the organizational level for data analysis. In order to perform the aggregation the name of the client / vendor organization is required. After the aggregation this information will be removed from the data set. All data is completely confidential. NO individual, team or company identifier information will be included in the analysis / report.

1. Respondent Information
   1.1. Please provide your name
   ____
   1.2. Please provide your designation
   ____
   1.3. Please provide the name of your organization
   ____
2. Information pertaining to the relationship.
   2.1. Please specify the type of process
   ____
   2.2. Please specify the client name
   ____
   2.3. Please specify the client industry
   ____
   2.4. Approximately, how long has this outsourcing arrangement been operational? __________
   2.5. Approximately, how long have you been part of this outsourcing arrangement? __________

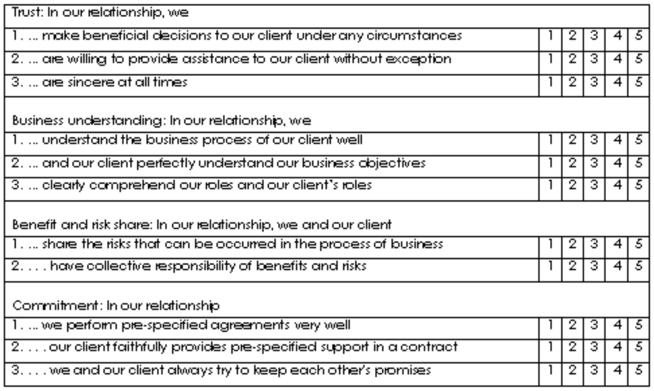
DEGREE OF CONTROL

In context of this outsourcing relationship ('queue'), please indicate the extent to which the client team uses the following means or mechanisms to influence the tasks, activities and day to day operations of the vendor team. **Please indicate your response on a scale of 1 to 5 where 1 indicates "rarely" and 5 indicates "quite a lot"**

FORMAL CONTROL					
Adherence to process documents	1	2	3	4	5
Incentives / Penalties	1	2	3	4	5
Performance metrics	1	2	3	4	5
Quarterly business reviews	1	2	3	4	5
Regular meetings or conference calls	1	2	3	4	5
Reports on transaction monitoring	1	2	3	4	5
INFORMAL CONTROL					
Client site visits	1	2	3	4	5
Giving free client merchandise	1	2	3	4	5
Providing recruitment guidelines	1	2	3	4	5
Providing training course material	1	2	3	4	5

PARTNERSHIP QUALITY

In context of this outsourcing relationship ('queue') please indicate the extent to which you agree or disagree with the following statements on a scale of 1 to 5, where 1 indicates "strongly disagree" and 5 indicates "strongly agree



Trust: In our relationship, we					
1. ... make beneficial decisions to our client under any circumstances	1	2	3	4	5
2. ... are willing to provide assistance to our client without exception	1	2	3	4	5
3. ... are sincere at all times	1	2	3	4	5
Business understanding: In our relationship, we					
1. ... understand the business process of our client well	1	2	3	4	5
2. ... and our client perfectly understand our business objectives	1	2	3	4	5
3. ... clearly comprehend our roles and our client's roles	1	2	3	4	5
Benefit and risk share: In our relationship, we and our client					
1. ... share the risks that can be occurred in the process of business	1	2	3	4	5
2. . . . have collective responsibility of benefits and risks	1	2	3	4	5
Commitment: In our relationship					
1. ... we perform pre-specified agreements very well	1	2	3	4	5
2. . . . our client faithfully provides pre-specified support in a contract	1	2	3	4	5
3. . . . we and our client always try to keep each other's promises	1	2	3	4	5

Thank you for your participation in this survey. In the space below, please provide any additional comments or feedback on this survey or any other aspect of this study that you may consider relevant.

APPENDIX B: CLIENT QUESTIONNAIRE

In this research study, the responses from the questionnaires will be aggregated at the organizational level for data analysis. In order to perform the aggregation the name of the client / vendor organization is required. After the aggregation this information will be removed from the data set. All data is completely confidential. NO individual, team or company identifier information will be included in the analysis / report.

1. Respondent Information
   1.1. Please provide your name
   ____
   1.2. Please provide your designation
   ____
   1.3. Please provide the name of your organization
   ____
   1.4. Please specify the industry your firm belongs to
   ____
2. Information pertaining to the relationship.
   2.1. Please specify the type of process outsourced
   ____
   2.2. Please specify the name of the outsource vendor
   ____
   2.3. Approximately, how long has this outsourcing arrangement been operational? ________
   2.4. Approximately, how long have you been part of this outsourcing arrangement? ________

Outsourcing Success:

In context of this outsourcing relationship ('queue') please indicate the extent to which you agree or disagree with the following statements on a scale of 1 to 5, where 1 indicates "strongly disagree" and 5 indicates "strongly agree.

This outsourcing relationship has produced beneficial economical results for the client (for instance reducing costs or capital expenditure)	1	2	3	4	5
This outsourcing relationship has produced beneficial process-related results for the client (for instance process clean-up, access to key technologies or skilled personnel)	1	2	3	4	5
This outsourcing relationship has produced beneficial business related results for the client (for instance enhanced focus on core competencies or new lines of business)	1	2	3	4	5
Overall, this outsourcing relationship has been valuable for the client	1	2	3	4	5
Overall, the client is satisfied with the results of this outsourcing relationship	1	2	3	4	5

Partnership Quality:

In context of this outsourcing relationship ('queue') please indicate the extent to which you agree or disagree with the following statements on a scale of 1 to 5, where 1 indicates "strongly disagree" and 5 indicates "strongly agree.

Trust: In our relationship,					
1. our service provider makes beneficial decisions to us under any circumstances	1	2	3	4	5
2. our service provider is willing to provide assistance to us without exception	1	2	3	4	5
3. our service provider is sincere at all times	1	2	3	4	5
Business understanding: In our relationship,					
1. we understand the business process of our service provider well	1	2	3	4	5
2. we and our service provider perfectly understand our business objectives	1	2	3	4	5
3. we clearly comprehend our roles and service provider's roles	1	2	3	4	5
Benefit and risk share: In our relationship,					
1. we and our service provider share the risks that can be occurred in the process of business	1	2	3	4	5
2. we and our service provider have collective responsibility of benefits and risks	1	2	3	4	5
Commitment: In our relationship					
1. our service provider performs pre-specified agreements very well	1	2	3	4	5
2. we faithfully provide pre-specified support in a contract	1	2	3	4	5
3. we and our service provider always try to keep each other's promises	1	2	3	4	5

Degree of Control:

In context of this BPO queue, please indicate the extent to which the following means or mechanisms is used to influence the tasks, activities and day to day operations of the vendor team. **Please indicate your response on a scale of 1 to 5 where 1 indicates "rarely" and 5 indicates "quite a lot"**

FORMAL CONTROL					
Adherence to process documents	1	2	3	4	5
Incentives / Penalties	1	2	3	4	5
Performance metrics	1	2	3	4	5
Quarterly business reviews	1	2	3	4	5
Regular meetings or conference calls	1	2	3	4	5
Reports on transaction monitoring	1	2	3	4	5
INFORMAL CONTROL					
Client site visits	1	2	3	4	5
Giving free client merchandise	1	2	3	4	5
Providing recruitment guidelines	1	2	3	4	5
Providing training course material	1	2	3	4	5

Thank you for your participation in this survey. In the space below, please provide any additional comments or feedback on this survey or any other aspect of this study that you may consider relevant.

____

____

____

____

____

____

Continued from page no. 12

[21] Igor E. Golovkin and Sushil J. Louis. Plasma x-ray spectra analysis using genetic algorithms. In Proceedings of the 1999 Genetic and Evolutionary Computing Conference (GECCO 1999), Orlando, Florida, 1999.

[22] Naik, Gautam. "Back to Darwin: In sunlight and cells, science seeks answers to high-tech puzzles." The Wall Street Journal, January 16, 1996.

[23] Geraldo Ribeiro Filho, Luiz Antonio Nogueira Lorena, “A Constructive Evolutionary Approach to School Timetabling”, 2001, Applications of Evolutionary Computing, EvoWorkshops2001: EvoCOP, EvoFlight, EvoIASP, EvoLearn, and EvoSTIM Proceedings

[24] “Genetic Approaches for Evolving Form in Musical Composition”, A. Ayesh and A. Hugill (UK), Proceeding 453, Artificial Intelligence and Applications 2/14/2005-2/16/2005 Innsbruck, Austria.

[25] Mahfoud, Sam and Ganesh Mani. "Financial forecasting using genetic algorithms", Applied Artificial Intelligence, vol.10, no.6, p.543-565 (1996).

[26] Coale, Kristi. "Darwin in a box." Wired News, July 14, 1997. Available online at http://www.wired.com/news/technology/0,1282,5152,00.html.

[27] Begley, Sharon and Gregory Beals. "Software au naturel." Newsweek, May 8, 1995, p.70.

[28] Garcia-Molina Hector, Ullman Jeffrey D, Widom Jennifer, “Database Systems – A Complete Book”, Pearson Education Singapore Pte. Ltd., 2004

[29] Ramkrishnan Raghu, Gehrke Johannes, “Database Management System”, 2nd edition, McGraw Hill, Singapore, 2000.

[30] Yu Clement T, Meng Weiyi, “Principles of Database Query Processing for Advanced Applications”, Morgan Kauffman SanFrancisco, 1998.Elmasri R, Navathe S B, “Fundamentals of Database System”, 4th edition, Pearson Education, 2004

[31] Özsu M Tamer, Valduriez, Ptrick, “Principles of Distributed Database Systems”, 2nd edition, Pearson Education Asia, 2001.

[32] Lanzelottel Rosana S G, Valdureiz Patrick, “Extending the search strategy in a Query Optimizer”, Proceedings of the 17th International Conference on Very Large Databases, Barcelona, September 1991.

[33] B. Kristin, M. C. Ferris, and Y. Ioannidis, "A genetic algorithm for database query optimization," Univ. Wisconsin, Madison, Tech. Rep. TR1004, 1991.

[34] Celko J, “Genetic Algorithms and Database Indexing”, Dr. Dobb's Journal, April 1993. [http://202.179.135.4/data/DDJ/articles/1993/9304/9304b/9304b.htm]

[35] Kratica Jozef, Ljubic Ivana, Tosic Dusan, “A Genetic Algorithm for index selection Problem”, In G. R. Raidl et al., editors, *Applications of Evolutionary Computing: EvoWorkshops2003*, volume 2611 of *LNCS*, pages 281-291, University of Essex, England, UK, 14-16 April 2003. Springer-Verlag.

[36] Martin Utesch, “Genetic Query Optimization in Database Systems”, PostgreSQL 7.1 Documentation for the Institute of Automatic Control at the University of Mining and Technology in Freiberg, Germany.

[37] Michael Steinbrunn, Guido Moerkotte, Alfons Kemper, “Heuristic and randomized optimization for the join ordering problem”, The VLDB Journal, 6(3), August 1997.

[38] Dembski, William “No Free Lunch: Why Specified Complexity Cannot Be Purchased Without Intelligence”, Rowman & Littlefield, 2002.

Modified Incremental Linear Discriminant Analysis for Face Recognition

R. K. Agrawal[1] and **Ashish Chaudhary**[2]

Abstract - ***Linear Discriminant analysis is a commonly used and valuable approach for feature extraction in face recognition. In this paper, we have proposed and investigated modified incremental Linear Discriminant Analysis (MILDA). We have compared the performance of proposed MILDA method against Pang et al ILDA in terms of classification accuracy, execution time and memory. It is found on the basis of experimental results with different face datasets that the proposed MILDA scheme is computationally efficient in terms of time and memory in comparison to batch method and Pang et al method. The experimental results also show that the classification accuracy due to MILDA, batch method and Pang et al are in complete agreement with each other.***

Index Terms - Statistical pattern recognition, Feature extraction, Face ecognition, Linear Discriminant Analysis

1. INTRODUCTION

Feature extraction is one of the important steps in pattern recognition which utilizes all the information of the given data to yield feature vector of the lower dimension and thereby eliminates redundant and irrelevant information. Commonly used feature extraction techniques are Principal Component Analysis (PCA) [1], Linear Discriminant Analysis LDA [2-7], Independent Components Analysis ICA [16] etc. Generally while applying these techniques for data classification it is assumed that complete dataset for training is available in advance and, learning is carried out in one batch. However in many real world applications such as pattern recognition and time series prediction we frequently come across situations where complete set of training samples is not available in advance, instead existing dataset keeps on changing with time. For example, in face recognition process human face undergoes facial variation due to different expressions (sad, happy, laughing face etc), lighting conditions, and make up, hairstyles etc. Hence, it is difficult to consider all facial variation when a human face is registered in a face recognition system, first time [8]. Similarly, in intrusion detection system, it is desirable to study the pattern behavior of intruder which can change slightly from its original behavior on account of incremental changes in data set [13]. Hence it is difficult to extract meaningful features only from previously available dataset. A straightforward approach in this situation is that we can collect data whenever new data are presented and then construct a provisional system by batch learning [9] over the collected data so far. However, such system will require large memory and high computational cost because the system would need to maintain a huge memory to store the data either previously learned, or newly presented, possibly without a limit. Moreover, the system will not be able to utilize the knowledge acquired in the past, even if the learning of most of the data is finished, and will repeat the learning from the beginning whenever one additional sample is presented.

To address such situation several eigenspace model [10–12] have been proposed. Hall, Marshall and Martin [10] proposed Incremental PCA (IPCA) based on the updating of covariance matrix through a residue estimating procedure. Agrawal and Karmeshu [11] proposed perturbation scheme for online learning of features based on incremental principal component analysis. Recently Pang, Ozawa and Kasabov [9] proposed an incremental linear discriminant analysis (ILDA). In this paper they have updated within class scatter matrix as new samples is added in terms of previously computed scatter matrix. This reduces the cost of computing updated scatter matrix. However the inverse of updated matrix S_W^* carried out for the computation of transformation matrix does not employ the previous knowledge of inverse of S_W. In this paper we investigate computationally more efficient scheme to compute the inverse of scatter matrix S_W^* which allows computation of dominant eigenvalues and eigenvector much more efficiently thereby increasing the performance of face recognition system.

The paper is organized as follows. Section 2 first provides a brief introduction of sequential ILDA proposed by Pang, Ozawa and Kasabov [9]. Following this, a modified incremental Linear Discriminant Analysis (MILDA) method is proposed in section 3. The performance of the proposed MILDA method in relation to batch method and Pang et al method is examined in terms of discriminability, computational time and memory in section 4. For this we have considered three publicly available face datasets [15]. The last section 5 contains conclusions.

2. SEQUENTIAL INCREMENTAL LDA

Suppose that initially we have a set of $\mathbf{N}$ $\mathbf{d}-$dimensional samples $\mathbf{x}_1, \mathbf{x}_2, \ldots\ldots\ldots \mathbf{x}_N$ belonging to $\mathbf{c}$ different classes with $\mathbf{N}_\mathbf{i}$ samples in the $\mathbf{i}^{\mathbf{th}}$ class. Then, the objective of LDA is to seek the direction $\mathbf{w}$ which maximizes the between-class scatter and minimizes the within class scatter of the projected images, such that the following criterion function [1]:

$$\mathbf{J}(\mathbf{w}) = \frac{\mathbf{w}^\mathbf{T}\mathbf{S}_\mathbf{B}\mathbf{w}}{\mathbf{w}^\mathbf{T}\mathbf{S}_\mathbf{W}\mathbf{w}} \qquad (1)$$

is maximized, where S_B and S_W are between-class scatter and within class scatter matrices and are defined as

[1,2] *School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi – 110067*
E-Mail: [1]rka@mail.jnu.ac.in, [2]ashish01jnu@gmail.com

$$\mathbf{S_B} = \sum_{1}^{c} \mathbf{N_i}\left(\mathbf{\bar{x}_i} - \mathbf{\bar{x}}\right)\left(\mathbf{\bar{x}_i} - \mathbf{\bar{x}}\right)^{\mathbf{T}} \qquad (2)$$

Where $\mathbf{\bar{x}}$ is d- dimensional sample mean of all the images and defined by

$$\mathbf{\bar{x}} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{1}^{n} \mathbf{x_k} \qquad (3)$$

and $\mathbf{\bar{x}_i}$ is i^{th} class mean given by

$$\mathbf{\bar{x}_i} = \frac{\mathbf{1}}{\mathbf{N_i}} \sum_{\mathbf{x_i \in C_i}} \mathbf{x_i} \qquad (4)$$

Within-class scatter matrix $\mathbf{S_W}$ is defined as

$$\mathbf{S_W} = \sum_{1}^{c} \mathbf{S_i} \qquad (5)$$

Where $\mathbf{S_i}$ is i^{th} -class scatter matrix

$$\mathbf{S_i} = \sum_{\mathbf{x \in c_i}} \left(\mathbf{x} - \mathbf{\bar{x}_i}\right)\left(\mathbf{x} - \mathbf{\bar{x}_i}\right)^{\mathbf{T}} \qquad (6)$$

To find the transformation matrix W, a generalized eigenvalue problem needs to be solved which is given by

$$\mathbf{S_B w} = \lambda \mathbf{S_w w} \qquad (7)$$

If S_w is non-singular then we have

$$\mathbf{S_W}^{-1}\mathbf{S_B w} = \lambda \mathbf{w} \qquad (8)$$

Hence in case of LDA the transformation matrix $\mathbf{W}$ is represented in terms of eigenvectors of matrix $\mathbf{U} = \mathbf{S_w}^{-1}\mathbf{S_B}$. It is obvious that the parameter needed for classification at any point of time are $\mathbf{S_W}, \mathbf{S_B}, \mathbf{\bar{x}}$ and $\mathbf{N}$. So discriminant eigenspace can be represented as $\Omega = \left(\mathbf{S_W}, \mathbf{S_B}, \mathbf{\bar{x}}, \mathbf{N}\right)$.

The traditional LDA works in a batch mode assuming that the whole dataset is given in advance and is trained in one batch only [9]. However, in a streaming environment, addition of any new samples will result in changes in original mean vector $\mathbf{\bar{x}}$, within class scatter matrix $\mathbf{S_W}$, as well as between-class distance matrix $\mathbf{S_B}$. Hence, the discriminant eigenspace model Ω needed to be updated. Pang, Ozawa and Kasabov [9] suggested incremental linear discriminant analysis (ILDA) for updating discriminant eigenspace Ω.

Pang et al [9] proposed that as a new sample $\mathbf{y}$ belonging to $\mathbf{k}^{\mathbf{th}}$ class is added to existing samples with mean $\mathbf{\bar{x}}$, within class scatter matrix S_W and the between class scatter matrix $\mathbf{S_B}$ then the new mean vector, the new between scatter matrix and the new within class scatter matrix are respectively given by $\mathbf{\bar{x}}^*$, S_W^* and $\mathbf{S_B^*}$ i.e.

$$\mathbf{\bar{x}}^* = \frac{\mathbf{N\bar{x} + y}}{\mathbf{N + 1}} \qquad (9)$$

If $\mathbf{k} = \mathbf{c} + \mathbf{1}$ i.e. the incoming sample belongs to a new class, then updated between-class scatter will be

$$\mathbf{S_B^*} = \sum_{\mathbf{i=1}}^{\mathbf{c+1}} \mathbf{N_i}^*\left(\mathbf{\bar{x}_i}^* - \mathbf{\bar{x}}^*\right)\left(\mathbf{\bar{x}_i}^* - \mathbf{\bar{x}}^*\right)^{\mathbf{T}} \qquad (10)$$

Where N_i^* is the number of samples in class i after addition of y. If $\mathbf{1 \le i \le c}$ then the updated matrix $\mathbf{S_B^*}$ is given by

$$\mathbf{S_B^*} = \sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{N_i}\, {}^*\left(\mathbf{\bar{x}_i}^* - \mathbf{\bar{x}}^*\right)\left(\mathbf{\bar{x}_i}^* - \mathbf{\bar{x}}^*\right)^{\mathbf{T}} \qquad (11)$$

Where $\mathbf{\bar{x}_i}^* = \left(\mathbf{1}/\left(\mathbf{N_i + 1}\right)\right)\left(\mathbf{N_i \bar{x}_i + y}\right)$ and $\mathbf{N_i^*} = \mathbf{N_i + 1}$, if $\mathbf{y}$ belongs to class i otherwise $\mathbf{\bar{x}_i}^* = \mathbf{\bar{x}_i}$ and $\mathbf{N_i}^* = \mathbf{N_i}$.

If y is a new class sample, which means $\mathbf{k}$ is the $\left(\mathbf{c+1}\right)^{\mathbf{th}}$ class, then the updated within class scatter matrix does not change:

$$S_W^* = \sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{S_i} + \mathbf{S_k} = \sum_{\mathbf{i=1}}^{\mathbf{c+1}} \mathbf{S_i} = \sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{S_i} = \mathbf{S_W} \qquad (12)$$

However if $1 \le i \le c$, then the updated S_W matrix is given by [9]

$$S_W^* = \sum_{\mathbf{i=1, i \ne k}}^{\mathbf{c}} \mathbf{S_i} + S_k^*$$

$$S_k^* = \mathbf{S_k} + \frac{\mathbf{N_k}}{\mathbf{N_k + 1}}\left(\mathbf{y} - \mathbf{\bar{x}_k}\right)\left(\mathbf{y} - \mathbf{\bar{x}_k}\right)^{\mathbf{T}} \qquad (13)$$

To determine transformation matrix $\mathbf{W}$, dominant eigenvectors of $\mathbf{U} = S_W^{*\,-1}\mathbf{S_B}\,^*$ is computed. However, this requires evaluation of inverse of matrix S_W^* i.e. $(S_W^*)^{-1}$ which is highly computational intensive operation. It would be useful to determine an alternative approach to compute inverse of $(S_W^*)^{-1}$ which reduces the cost of computation without decreasing its accuracy.

3. ODIFIED INCREMENTAL LINEAR DISCRIMINANT ANALYSIS (MILDA)

It will be noteworthy if we are able to calculate $(S_W^*)^{-1}$ in terms of the previously calculated S_w^{-1}, thereby decreasing the cost of computation. Equation (13) can also be rewritten as

$$S_W^* = \mathbf{S_W} + \frac{\mathbf{N_k}}{\mathbf{N_k + 1}}\left(\mathbf{y} - \mathbf{\bar{x}_k}\right)\left(\mathbf{y} - \mathbf{\bar{x}_k}\right)^{\mathbf{T}} \qquad (14)$$

According to Woodbury Formula [14]: If A is a matrix of dimension $n \times n$ and U and V are vectors of size n then

$$\left(\mathbf{A}+\mathbf{U}\mathbf{V}^{\mathbf{T}}\right)^{-1}=\mathbf{A}^{-1}-\frac{\mathbf{A}^{-1}\mathbf{U}\mathbf{V}^{\mathbf{T}}\mathbf{A}^{-1}}{1+\mathbf{V}^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{U}} \qquad (15)$$

Equation (15) allows computing inverse of a perturbative matrix in terms of a given matrix A and change to the given matrix A.

Using (14) and (16), we get

$$(S_W^*)^{-1}=\mathbf{S_W}^{-1}-\frac{\mathbf{S_W}^{-1}\mathbf{U}\mathbf{V}^{\mathbf{T}}\mathbf{S_W}^{-1}}{1+\mathbf{V}^{\mathbf{T}}\mathbf{S_w}^{-1}\mathbf{U}} \qquad (16)$$

Where $\mathbf{U}=\frac{\mathbf{N_k}}{\mathbf{N_k}+\mathbf{1}}\left(\mathbf{y}-\overline{\mathbf{x}}_{\mathbf{k}}\right)$ and $\mathbf{V}=\left(\mathbf{y}-\overline{\mathbf{x}}_{\mathbf{k}}\right)$.

The computation of inverse of a matrix of size $n\times n$ requires $O(n^3)$ time. However, inverse of a matrix S_W^* in terms of inverse of matrix S_w can be computed in $o(n^2)$ time thereby decreasing the cost of computation. Hence, it will be more appropriate to represent discriminant eigenspace by $\mathbf{\Phi}=\left(\mathbf{S_W}^{-1},\mathbf{S_B},\overline{\mathbf{x}},N\right)$ which allows updating eigenspace when a new sample is considered in addition to existing samples.

The outline of the procedure based on MILDA is given below:

MILDA Algorithm

Input: [X₁, X₂… Xₙ]

```
1. Compute $\overline{\mathbf{x}}$, $\mathbf{S}_B$, $\mathbf{S}_W$
2. Compute $\mathbf{S}_W{}^{-1}$
3. For each tuple y do the following:
     If y belongs to new class then
         Compute $\mathbf{S}_B^*$ using equation (10)
         $(S_W^*)^{-1}=\mathbf{S_W}^{-1}$
     Else
         Compute $\mathbf{S}_B^*$ using equation (11)
         $(S_W^*)^{-1}=\mathbf{S_W}^{-1}-\frac{\mathbf{S_W}^{-1}\mathbf{U}\mathbf{V}^{\mathbf{T}}\mathbf{S_W}^{-1}}{1+\mathbf{V}^{\mathbf{T}}\mathbf{S_w}^{-1}\mathbf{U}}$
     End
4. Compute (c-1) dominant eigenvectors of $(S_W^*)^{-1}\mathbf{S}_B^*$ i.e. $[\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_{(c-1)}]$
```

Output: W =[e₁, e₂, …, e₍c-1₎]

4. EXPERIMENTAL SETUP AND RESULTS

In this section, we have examined the efficiency and accuracy of our modified approach for updating discriminant eigenspace i.e. $\mathbf{\Omega}=(\mathbf{S}_{\mathbf{W}}^{\mathbf{-1}},\mathbf{S_B},\overline{\mathbf{x}},\mathbf{N})$ against Sequential ILDA $\Omega=(S_W,\mathbf{S_B},\overline{\mathbf{x}},\mathbf{N})$ proposed by Pang et al. The modified approach is evaluated in terms of discriminability, execution time (time taken to update eigenspace) and memory usage against Sequential ILDA. For all experiments matlab code running on a PC with Intel Pentium 4 2.8 GHz CPU and 256-Mb RAM is used.

Extensive experiments are carried out on three publicly available databases [15]: Yale, ORL, and JAFFE to check the efficacy of the proposed MILDA. The ORL database [15] consists of 40 different individuals with 10 images for each individual. All the images are taken against a dark homogeneous background and the subjects are in up-right, frontal position (with tolerance for some side movement). The images from ORL database were cropped from 112 × 92 to 75 × 50 in our experiment. The Yale database [15] consists of 165 images, which are made up of 16 different individuals with 11 images for each individual. The size of images is changed from 320 × 243 to 100 × 102 in our experiment. The JAFFE database [15] comprises 10 Japanese females. Each person has seven facial expressions: "happy," "sad," "surprise," "angry," "disgust," "fearful," and "neutral." There are three or four images for each facial expression of each person. The images from JAFFE database were cropped from 256× 256 to 128 × 128 in our experiment.

For every test, first we constructed an initial feature space using 20% of the total samples in which at least one image from each class is ensured to be included. For carrying out incremental learning one sample is chosen randomly from the remaining training samples. For incremental learning, we first encode features by projecting data presented in terms of updated eigenspace. We have used K-nearest neighbor classifier (K-1) [17] in our experiments. The "leave-one-out" strategy is adopted for testing and training. We found that proposed MILDA method can classify data with same accuracy as batch method and Pang et al method for all the three face datasets.

The computational gain of proposed MILDA scheme for calculating discriminant eigenspace $((S_W^*)^{-1},S_B^*)$ as compared to Pang et al method and batch method is shown in Figure 1 for the ORL database. It can be observed that the difference between the execution time in batch method and both incremental methods (Pang et al method and MILDA) is quite significant. It can also be observed that the difference between the execution time in proposed MILDA scheme and Pang et al method is not significant for small set of samples but becomes pronounced as sample size increases.

Figures 2-3 shows the variations in execution time to compute discriminant eigenspace $((S_W^*)^{-1},S_B^*)$ with number of samples for MILDA and Pang et al method for ORL and JAFFE face datasets. Figures 2-3 show that the proposed MILDA scheme requires less computation time in comparison to Pang et al method. It can also be observed that the difference in execution time to compute eigenspace $((S_W^*)^{-1},S_B^*)$ is more significant when sample size is large. Experimental results on Yale dataset also show that the proposed MILDA scheme outperforms Pang et al method in terms of computation time.

We have also estimated the amount of memory required by Pang et al method and MILDA method for incremental linear Discriminant analysis for face datasets. The results are shown in Figures 5-7. It can be observed that memory requirement is more in Pang et al method in comparison to MILDA method when sample size is large for all the three face datasets.

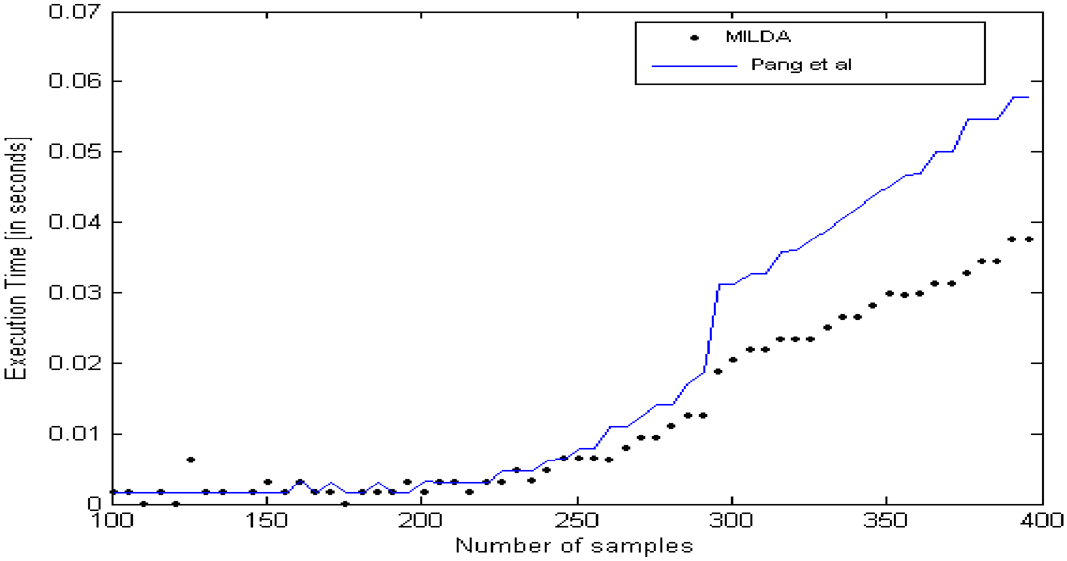


Figure 1: Variation in Execution time for Batch method, Pang et al method and MILDA method when new features are added for ORL dataset

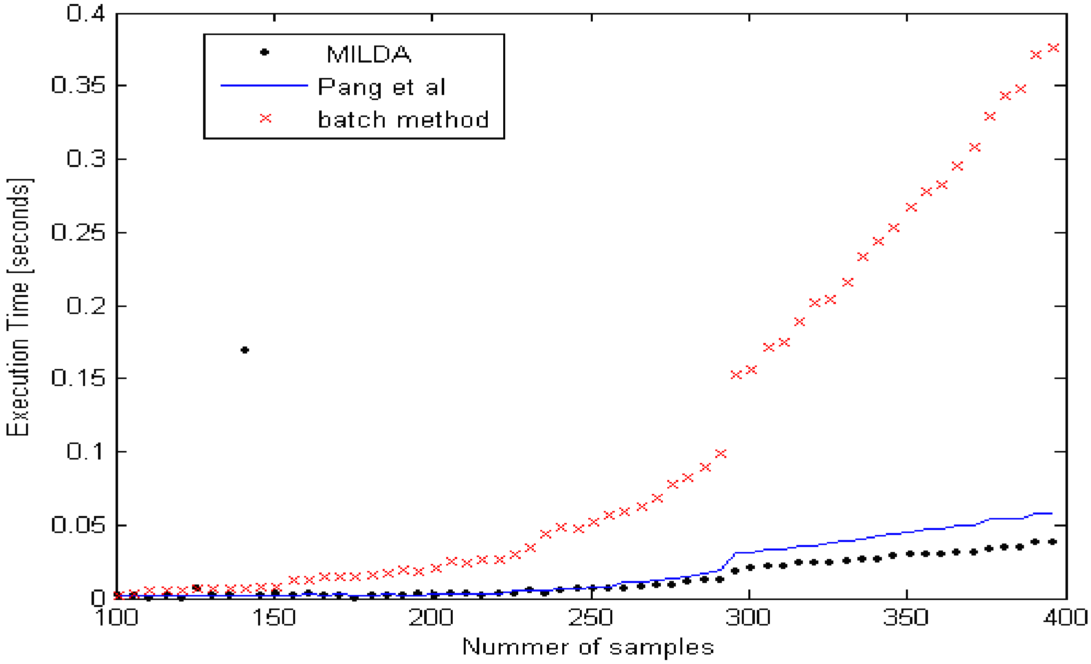


Figure 2: Variation in Execution time for Pang etal method and MILDA method when new features are added for ORL datas

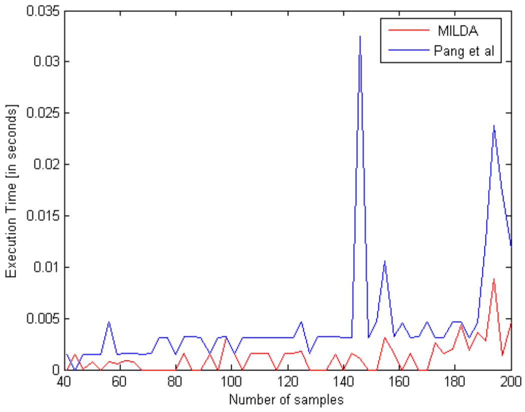


Figure 3: Variation in Execution time for Pang et al method and MILDA method when new features are added for JAFFE dataset

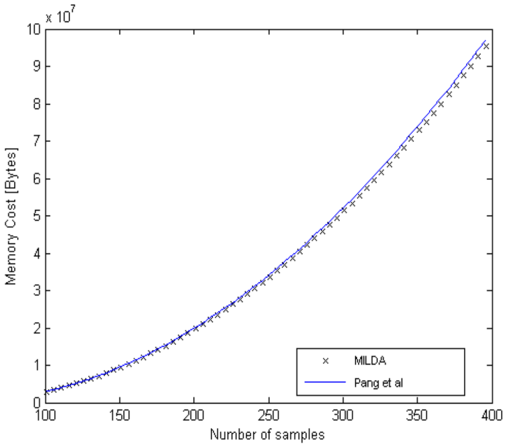


Figure 5: Variation in Memory usage for Pang et al method and MILDA method when new features are added for ORL dataset

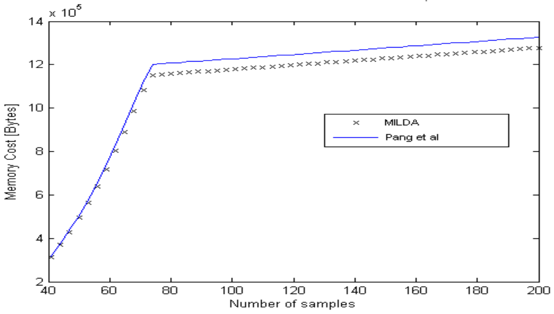


Figure 6: Variation in Memory usage for Pang et al method and MILDA method when new features are added for JAFFE dataset

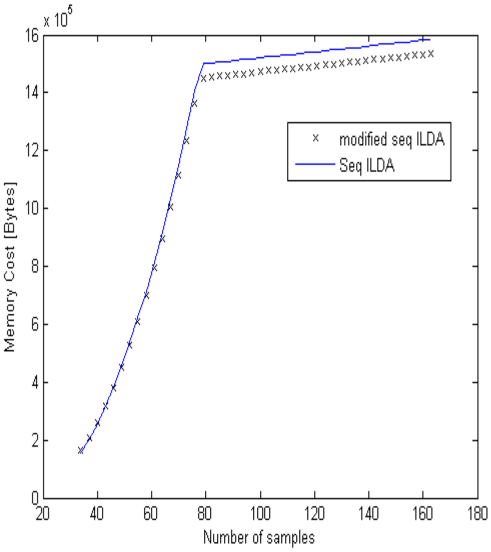


Figure 7: Variation in Memory usage for Pang et al method and MILDA method when new features are added for Yale dataset

5. CONCLUSION

In this paper, we have proposed and investigated modified incremental Linear Discriminant Analysis. Results from this study suggest that the modified incremental Linear Discriminant Analysis is computationally more efficient in comparison to batch method and Pang et al method. This is due to the fact that the batch method and Pang et al method involve intensive matrix operations. The time complexity of inverse of a matrix of size $n \times n$ requires O(n^3) whereas the proposed scheme requires O(n^2). The performance is evaluated in terms of (a) Discriminability, (b) Time required to carry out inverse of matrix (b) Memory requirement. Experimental results with different face datasets show that the proposed scheme is computationally more efficient in terms of time and memory in comparison to batch method and Pang et al method. These investigations suggest that the proposed scheme may be useful for online face detection systems where both memory and computation time is of utmost importance.

REFERENCES

[1]. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs.Fisherfaces: Recognition using class specific linear projection, ”*IEEE Trans. Pattern Anal Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[2]. D. L. Swets and J.Weng, “Using discriminant eigenfeatures for image retrieval,” *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 18, no. 8, pp. 831–836, Aug. 1996.

[3]. L. Chen, H. M. Liao, M. Ko, J. Lin, and G. Yu, “A new LDA-based face recognition system which can solve the small sample size problem,” *Pattern Recognit.*, vol. 33, pp. 1713–1726, Oct. 2000.

[4]. A. M. Martlnez and A. C. Kak, “PCA versus LDA,” *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.

[5]. H.Yu and J. Yang, “A direct LDA algorithm for high-dimensional data with application to face recognition,” *Pattern Recognit.*, vol. 34, pp. 2067–2070, 2001.

[6]. C. Liu and H. Wechsler,“Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition,” *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[7]. X. Wang and X. Tang,“A unified framework for subspace face recognition,” *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 26, no. 9, pp. 1222–1228, Sep. 2004.

[8]. S. Ozawa, S. Pang, N. Kasabov, A modified incremental principal component analysis for online learning of feature space and classifier, in C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.), PRICAI : Trends in Artificial Intelligence LNAI, Springer-Verlag (2004) 231-240.

[9]. S. Pang, S. Ozawa, and N. Kasabov, “, Incremental Linear Discriminant Analysis for Classification of Data Streams” IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 35, NO. 5, OCTOBER 2005

[10]. P. M. Hall, A. D. Marshall, R. R. Martin, “Incremental eigenanalysis for classification,” in *Proc. Brit. Machine Vision Conf.*, vol. 1, pp. 286–295.

[11]. R. K Agrawal and Karmeshu,’’ Perturbation scheme for online learning of features ,Incremental Principal Component Analysis”. Pattern Recognition, vol 41, no,. 5, pp. 1452-1460, (2008).

[12]. S. Chandrasekaran, B. S. Manjunath, Y. F. Wang, J. Winkler, and H. Zhang, “An eigenspace update algorithm for image analysis,” *Graphical Models Image Process.*, vol. 59, no. 5, pp. 321–332, Sep. 1997.

[13]. S. Mukkamala, A. H. Sung, Artificial intelligent techniques for intrusion detection, IEEE International Conference on Systems, Man and Cybernatics **2** (2003) 1266-1271.

[14]. Gene H. Golub, Charles F. Van Loan, Matrix Computations, 3rd Edition, The John Hopkins University Press Baltimore, 1996

[15]. http:// www.face-rec.com (Face Recognition Homepage)

[16]. M.SBartlett and T.JSejowski,” Independent Component of face images: A representation for Face Recognition” in Proc.4th Ann.Jount Symp. Neural Comput. Pasadena,CA,may 17,1997.

[17]. R. O. Duda, P. E. Hart and D. G. Stork, Pattern classification, Second Edition,Wiley, 2001

Process Centric Development to Improve Quality of Service (QoS) in Building Distributed Applications

K. Krishna Mohan[1], **A.Srividya**[2] , **A. K. Verma**[3] and **Ravi Kumar Gedela**[4]

Abstract - In a competitive business landscape, large organizations such as insurance companies, banks etc are under high pressure to innovate, improvise and differentiate their products and services while continuing to reduce the time-to market for new product introductions. For example, in banks operating multiple lines of businesses, generating a Single view of the customer is a critical from different perspectives due the systems developer over a period of time and existence of disconnected systems with an enterprise. Therefore to increase revenues and cost optimization, it is important build enterprise systems more closely with the business requirements with re-use of existing systems. While building distributed based applications, it is important to take into account the proven processes like Rational Unified Process (RUP) to mitigate the risks and increase the reliability of the system. Experience in developing applications in Java Enterprise Edition (JEE) with customized RUP is presented in this paper. RUP is adopted into an onsite-offshore development model along with ISO 9001and SEI CMM Level 5 standards. This paper provides RUP approach to achieve increased reliability with higher productivity with lower defect density and competitiveness through cost effective custom software solutions.

Index Terms - Rational Unified Process, Java Enterprise Edition (JEE), Phases, Disciplines, Reliability, Testing, Artifacts, Use Case and Model, metrics, Offshore Development Center (ODC).

1. INTRODUCTION

development has expanded rapidly in recent years and has brought a wake of changes that impact application development projects [1]. The adoption of a new process for delivery excellence within an organization is critical to meet time-to-market conditions and it is a significant undertaking. It requires careful customization to match the organization culture, accommodate any existing procedures, and obtain buy-in among the key stakeholders and users of the process. Many organizations have initiated a program to standardize the software development process using rational tools and RUP. A natural extension of adoption of RUP would be extending the same to offshore development to reduce the total cost of ownership (TCO) with improved reliability with higher productivity. However, with the comprehensive nature of RUP comes significant complexity regarding the process steps and types of artifacts produced at each step. The authors therefore created this paper to cover the elements of a RUP-based development process that are vital to a successful development projects.

2. BACKGROUND AND MOTIVATION

National Research Council Canada [2] and several other organizations reveal that the process oriented development is necessary to improve the reliability and productivity and decrease the cost, thereby increase the operational efficiency. This paper provides an approach to adopt RUP in building applications and focusing on various areas of software development. It discusses the approach citing examples of the work done by authors in two areas: Requirements Gathering/Modeling and testing (Manual and automation) phases.

A Proof-of–Concept (PoC) with financial domain application is developed and tested to address the RUP approach to demonstrate the benefits. The PoC was developed in SOA - J2EE platform with RUP and adoption of RUP for several projects the results obtained and conclusions are being shared in this paper.

3. GENERIC ENGAGEMENT MODEL (ONSITE-OFFSHORE MODEL WITHOUT RUP)

Based on the working experience of the authors it is well accepted that most of the service organizations follow the onsite-offshore model to reduce the development and maintenance costs. Development in India or other low cost centers (here after referred as offshore development) teams are effectively leveraged the onsite-offshore model to provide value based services. This model is true for most of the North American and European companies/customers.

Most of the organizations successful in following Iterative Incremental development methodology in order to meet the customer requirements through an early and continuous delivery of software to end users at regular intervals. The lifecycle in iterative development is composed of several iterations in sequence. Each iteration is a self-contained mini-release composed of activities such as requirements, analysis, design, development, and testing. The final system is built by adding and releasing new features in each iteration. Every iteration ends up with:

1. The delivery of stable release
2. A visual model of emerging product
3. Lessons learned to incorporate into next iteration
4. Customer sign-off on implementation of requirements

[1] *RS, Reliability Engineering Group, Department of Electrical Engineering, Indian Institute of Technology Bombay*
[2, 3] *Professor, Reliability Engineering Group, Department of Electrical Engineering, Indian Institute of Technology,*
[4] *HCU – SAP, Satyam Computer Services Limited, Hyderabad*
E-Mail: [1]kkm@ee.iitb.ac.in, [2]asvidya@ee.iitb.ac.in,
[3]akv@ee.iitb.ac.in and [4]Ravikumar_Gedela@satyam.com

The diagram in figure 1 depicts the iterative and incremental development process that is being followed currently.

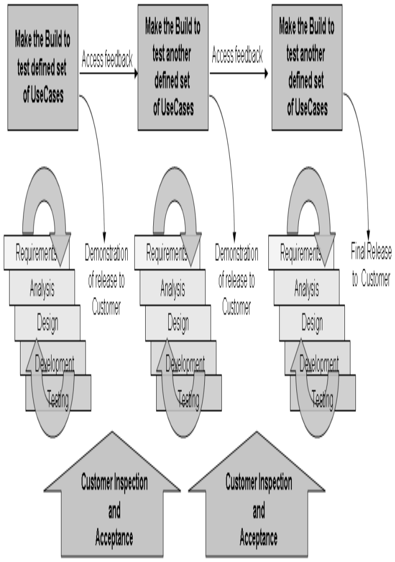

Make the Build to test defined set of UseCases
Access feedback
Make the Build to test another defined set of UseCases
Access feedback
Make the Build to test another defined set of UseCases
Requirements
Analysis
Design
Development
Testing
Demonstration of release to Customer
Demonstration of release to Customer
Final Release to Customer
Customer Inspection and Acceptance
Customer Inspection and Acceptance

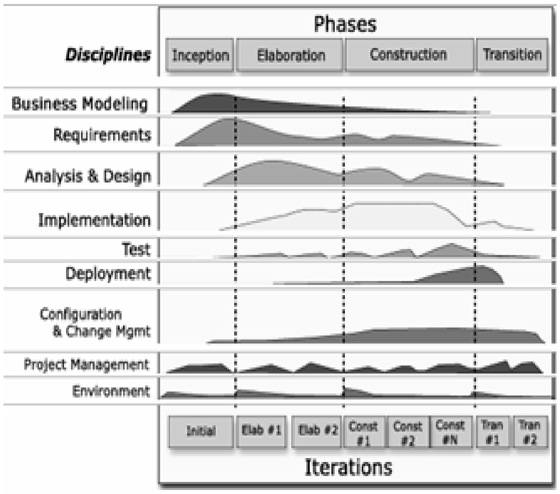
Figure 1: Generic Engagement Model

The key characteristics of such developments are:

1. The development is framework driven. It is necessary that the onsite and offshore team have understood the framework on which the development is based. The framework is in terms of the base architecture, definition of common components, clear understanding of the code and design practices and so on.
2. The onsite team manages the scope. Any new features are discussed and negotiated with the customer. They are put on wait till next release if they are impacting the schedule.
3. Involvement of business users in the development at early stages is key. From our experience, such projects are successful with full involvement of the business people. The onsite team's one of the main responsibility is to engage with business for clarifications of requirements and JAD sessions
4. Planning is done iteratively. There is a possibility that such development may fall in the "death by planning trap". The plan is iterative, the milestones are fixed and they are clearly communicated to all stakeholders. A high-level, coarse-grained view of the project is developed during the initial iteration. It shows the total number of planned iterations across all the phases, and key milestone dates for each of these iterations. A fine grained plan is made for each iteration.
5. There is lot of emphasis on separate test team. This team works with the development team and continuous testing is done during the development.
6. The configuration management plan is clearly laid out. The activities are more frequent. Generally, in such projects, the offshore team delivers code drops once in three days. The onsite team continuously integrates the code drops.
7. The product works from day one. The functionality may be very minimal but the product is built very frequently.
8. Effective risk management plan. The risks are monitored on weekly basis and are communicated to the customer.

4. RATIONAL UNIFIED PROCESS

RUP is a comprehensive framework; it is a more or less complete set of process elements that has to be tailored to each case as no project needs the complete set of elements. IBM Rational has in fact done some of the tailoring of the original Unified Process by the development of RUP [3]. The Rational Unified Process® [4,5] or RUP® product is a software engineering process. It provides a disciplined approach to assigning tasks and responsibilities within a development organization. Its goal is to ensure the production of high-quality software that meets the needs of its end users within a predictable schedule and budget.



Source: IBM Corporation
Figure 2: Rational Unified Process

The figure 2 illustrates the overall architecture of the RUP, which has two dimensions:

1. The horizontal axis represents time and shows the lifecycle aspects of the process as it unfolds. This first dimension illustrates the dynamic aspect of the process as it's enacted and is expressed in terms of phases, iterations, and milestones.
2. The vertical axis represents disciplines that logically group activities by nature. This second dimension portrays the static aspect of the process-how it's described in terms of process components, disciplines, activities, workflows, artifacts, and roles.

The graph shows how the emphasis varies over time. For example, in early iterations you spend more time on requirements; in later iterations you spend more time on implementation.

5. CUSTOMIZED RUP FOR ONSITE-OFFSHORE MODEL DELIVERY

As described above, RUP is iterative, use case based, Architecture driven process, it can be customized for an organization. The following Figure 3 customized for a leading bank in North America and which is operational.

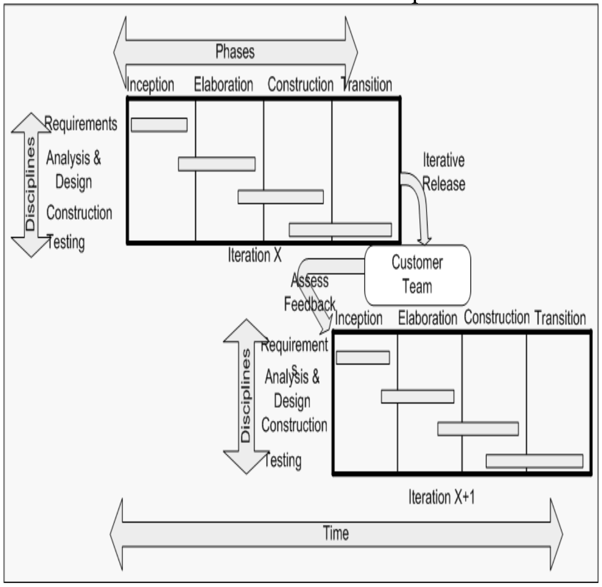
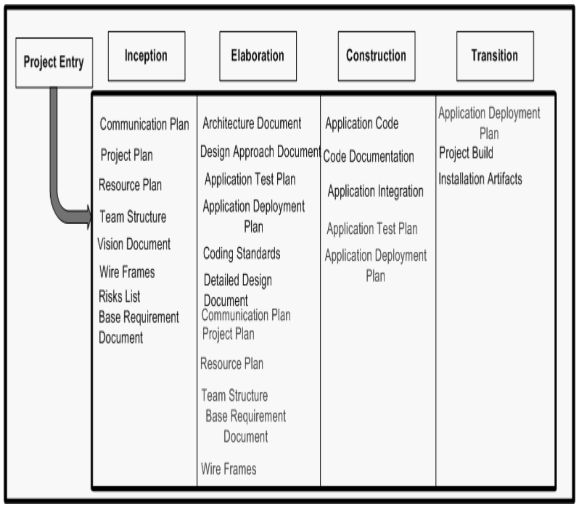


Figure 3: Customized RUP

Four phases are defined for a project lifecycle, which proceed in order:

- **Inception**: the beginning phase of the project with priorities on achieving concurrence among all stakeholders on the lifecycle objectives for the project.
- **Elaboration**: focuses on finalizing the system and software architecture and requirements
- **Construction**: building phase of the project that implements what's been laid out during the elaboration phase to produce an alpha-tested stable system
- **Transition**: final phase of the project that makes the system production ready and prepares the user community for use

The following diagram presents the RUP phase-end deliverables in an application environment. Note that these do not constitute the full set of artifacts normally produced during a project that uses RUP. Based on the previous engagements experience, authors suggests ensuring a reliable and repeatable development with the artifacts mentioned in Figure 4 at high level.



Inception	Elaboration	Construction	Transition
Communication Plan Project Plan Resource Plan Team Structure Vision Document Wire Frames Risks List Base Requirement Document	Architecture Document Design Approach Document Application Test Plan Application Deployment Plan Coding Standards Detailed Design Document Communication Plan Project Plan Resource Plan Team Structure Base Requirement Document Wire Frames	Application Code Code Documentation Application Integration Application Test Plan Application Deployment Plan	Application Deployment Plan Project Build Installation Artifacts

Figure 4: RUP Phase-end deliverables

As described in the previous section, key activities are carried in different phases (Inception, Elaboration, Construction and Transition) of RUP as depicted in the above diagram. During the project delivery, offshore is viewed as an extended development facility. To leverage the onsite-offshore model and effective delivery efficiency, the set of activities are distributed in between onsite and offshore facilities.

During the Inception phase, the project vision is established based on the requirements from business users and deliver Vision document. The next key activity of constructing the use cases from the requirements that are captured in the Inception phase is carried out in the Elaboration phase. In this phase, the following activities are carried out.

1. Complete set of use cases (UC) will be planned against various iterations and deliver Use-Case document with Iteration Plan.
2. High level design and Low Level Design will be generated as indicated in the Fig 4.

To leverage the offshore felicity effectively, key activities are being carried out from the elaboration iteration (X+1) and construction (X) will be carried out simultaneously. The (X+1) iteration of Elaboration phase will be carried out at onsite and the X iteration of Construction will be carried out at Offshore. After completion of the construction iteration X, the transition phase will be initiated. During the Transition phase, Integration System Testing (IST) execution will be carried out. After the completion of initial iterations of Transition phase, (X+1) iteration of Construction phase would start. This phase will be followed by (X+1) iteration of Transition phase. After completion of planned iterations, Quality Assurance Test (QAT) will be conducted in an integrated environment (which labeled as Transition III in the diagram Fig 3). After meeting the required QAT criteria, the project will be deployed/ transitioned to production.

6. CASE STUDY

This section explains the work done in building the Proof-of-Concept of a financial services application.

To narrate the simplified version of the application use case diagram is depicted. Use cases drive the whole development process. With each iteration, Use cases drive the work in analysis, design, implementation and test [6, 7]. This paper addresses the requirements phase and testing phase as the project is completely developed and currently in production and does not addresses the other phases in detail. The use case of the application is shown in the figure 5.

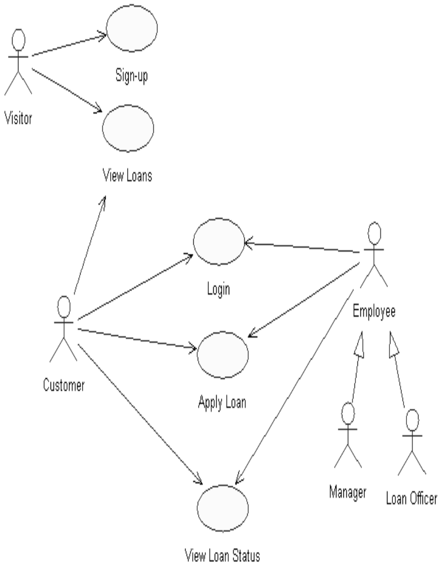


Figure 5: Use Cases of the application, few cases are considered for PoC

6.1 Logical Architecture of Credit Management Application with Design Patterns

The implementation model of the PoC with proven Design Patterns is as follows: The thin-client application users uses the browser to access the application. The request from the browser passes, on (over HTTP) to the Presentation Layer which implements the MVC design pattern. The industry wide proven open source framework 'struts' used to realize the presentation layer.

The Business Layer implements business processes for different modules as Java Objects and Enterprise Java Beans (EJB) which encapsulates the business logic. The Data Layer implements the Data Access Object (DAO) pattern. The data access components encapsulate the data sources from the business layer as shown in Figure 6.

6.2 Deployment of the Credit Management Application

The deployment of the Credit Management Application is as follows in the figure 7.

The application is developed in Java/J2EE running on Windows XP server and database is Oracle 9i. The IBM HTTP Server, IBM WebSphere Application Server and the Oracle Database were running on the on different systems system with Clustering. The clustering mechanism was chosen to demonstrate the Failover in the event if any WebSphere Application Server is down. The following Tables 2, 3, 4, 5, 6, 7 and 8 represent the application behavior at runtime.

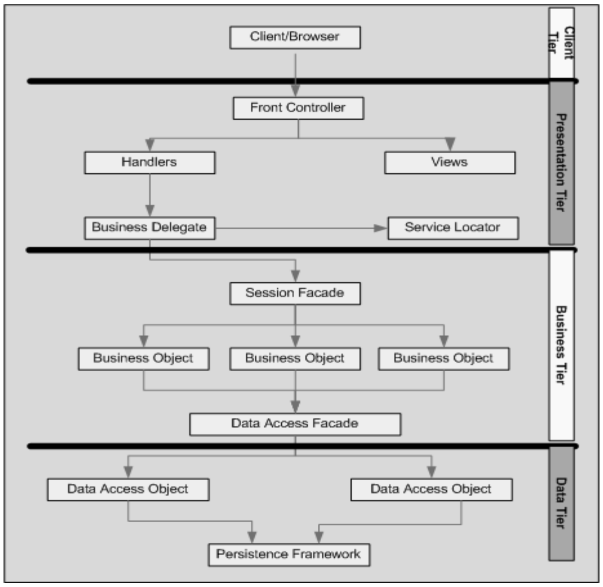


Figure 6: Logical Architecture of Trader Tax monitor with Design Patterns

Components and frameworks selected for J2EE version of PoC are provided in Table 1

Layer	Presentation	Business	Data
Technology Options	Jakarta Struts 1.2	Enterprise Java Beans (EJB) and POJOs	Hibernate

Table 1: J2EE selection

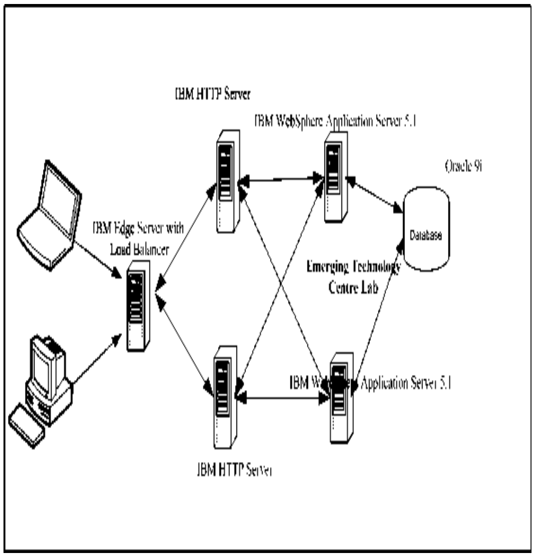

IBM HTTP Server
IBM WebSphere Application Server 5.1
Oracle 9i
Database
IBM Edge Server with Load Balancer
Emerging Technology Centre Lab
IBM W... Application Server 5.1
IBM HTTP Server

Figure 7: Deployment of the Credit Management Application

6.3 Database Design

CMS_USER_DETAILS			
Column Name	**Type**	**Null?**	**Constraint**
USER_ID	VARCHAR2(20)	NOT NULL	PRIMARY KEY
FIRST_NAME	VARCHAR2(20)	NOT NULL	
LAST_NAME	VARCHAR2(20)	NOT NULL	
STREET1	VARCHAR2(20)	NOT NULL	
STREET2	VARCHAR2(20)		
DISTRICT	VARCHAR2(20)	NOT NULL	
STATE	VARCHAR2(20)	NOT NULL	
COUNTRY	VARCHAR2(20)	NOT NULL	
EMAIL	VARCHAR2(20)		
PHONE	VARCHAR2(20)		
PIN	VARCHAR2(20)	NOT NULL	

Table 2

CMS_LOGIN			
Column Name	**Type**	**Null?**	**Constraint**
USER_ID	VARCHAR2(20)	NOT NULL	REFERENCES CMS_USER_DETAILS
PASSWORD	VARCHAR2(20)		
ROLE	VARCHAR2(20)		

Table 3

CMS_LOAN_DOCUMENTS			
Column Name	**Type**	**Null?**	**Constraint**
DOCUMENT_ID	NUMBER	NOT NULL	PRIMARY KEY
DOCUMENT_NAME	VARCHAR2(20)		
DOCUMENT_DESCRIPTION	VARCHAR2(20)		

Table 4

CMS_LOAN_TYPE			
Column Name	**Type**	**Null?**	**Constraint**
LOAN_TYPE_ID	NUMBER	NOT NULL	PRIMARY KEY
LOAN_TYPE	VARCHAR2(20)		
LOAN_DESCRIPTION	VARCHAR2(20)		

Table 5

CMS_LOAN_TYPE_DOCUMENTS			
Column Name	**Type**	**Null?**	**Constraint**
LOAN_TYPE_ID	NUMBER		REFERENCES CMS_LOAN_TYPE
DOCUMENT_ID	NUMBER		REFERENCES CMS_LOAN_DOCUMENTS

Table 6

CMS_LOAN_APPLICATION			
Column Name	**Type**	**Null?**	**Constraint**
LOAN_ID	VARCHAR2(20)	NOT NULL	PRIMARY KEY
LOAN_TYPE_ID	VARCHAR2(20)		REFERENCES CMS_LOAN_TYPE
USER_ID	VARCHAR2(20)		REFERENCESUSER_DETAILS
LOAN_AMOUNT	VARCHAR2(20)	NOT NULL	
TENURE	VARCHAR2(20)	NOT NULL	
APP_DATE	VARCHAR2(20)	NOT NULL	
STATUS	VARCHAR2(20)		
OFFICER_APPROVAL_DATE	VARCHAR2(20)		
OFFICER_APPROVAL_REMARKS	VARCHAR2(20)		
OFFICER_SUGGESTED	VARCHAR2(20		

_AMOUNT	)		
MANAGER_APPROVAL_DATE	VARCHAR2(20)		
MANAGER_APPROVAL_REMARKS	VARCHAR2(20)		
MANAGER_SUGGESTED_AMOUNT	VARCHAR2(20)		

Table 7

CMS_LOAN_TYPE			
Column Name	**Type**	**Null?**	**Constraint**
LOAN_TYPE_ID	NUMBER		REFERENCES CMS_LOAN_TYPE
DOCUMENT_ID	NUMBER		REFERENCES CMS_LOAN_DOCUMENTS
STATUS	VARCHAR(20)		

Table 8

6.4 Sequence diagrams

The following figures 8 – 11 are sequence diagrams for sign up, Login, Sign-out, Loan Application are shown below.

Sequence Diagram for Sign Up

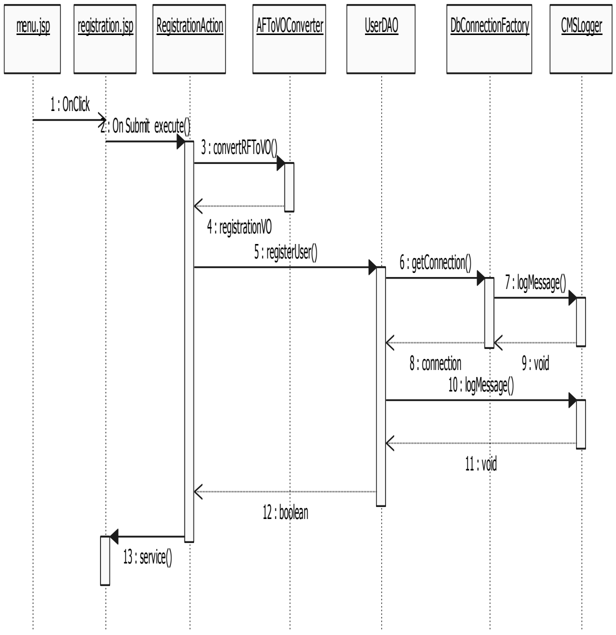


Figure 8: Sequence Diagram for Sign Up Login Sequence Diagram

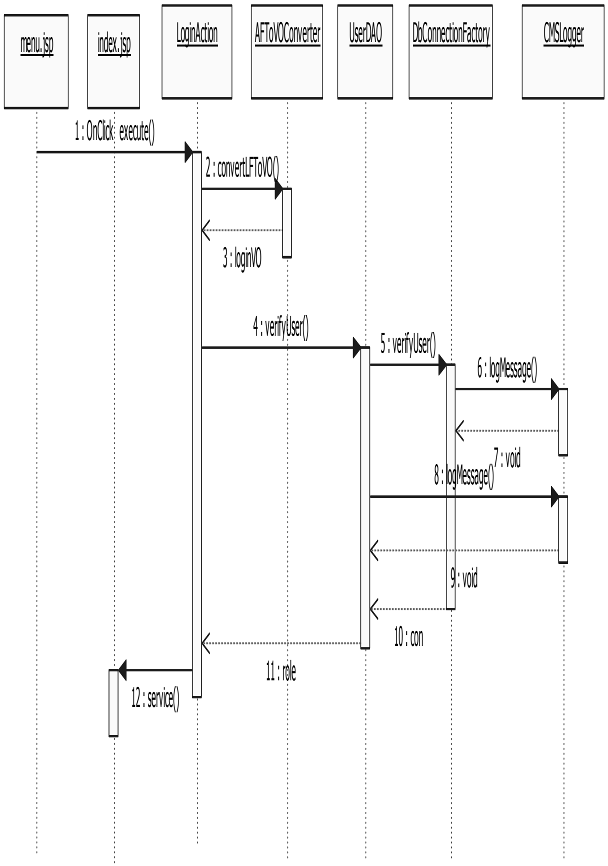


Figure 9: Sequence Diagram for Login

Sign-out

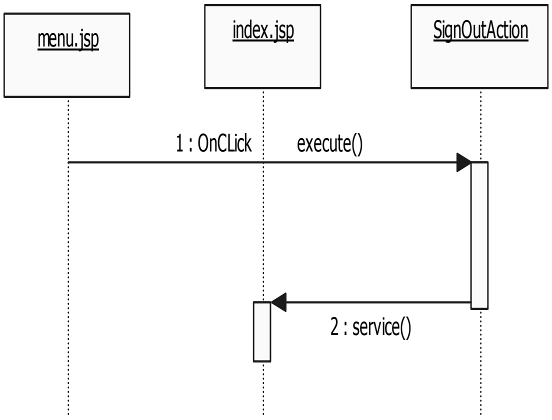


Figure 10: Sequence Diagram for Sign-out Loan Application

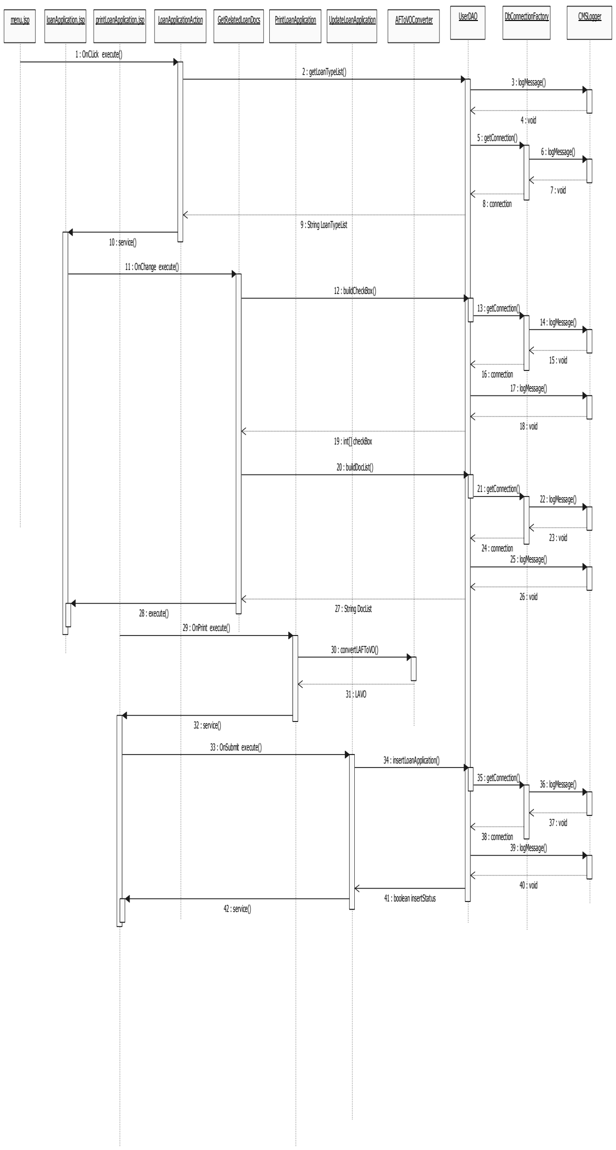


Figure 11: Sequence Diagram for Loan Application

7. REQUIREMENTS FLOW – RUP

The PoC is developed in Java/J2EE with strong object oriented principles. The authors suggest the following requirement flow to capture requirements.

As depicted by the Figure 12, the following major activities are executed for the requirements discipline

- **Analyze the Problem**: identify problem to be solved, system boundaries and constraints, and stakeholders.
- **Understand Stakeholder Needs**: elicit stakeholder requests or "wish list"
- **Identify Actors:** identify people, applications that will interact with the system
- **Find Use Case**: find system processes or set of behavior that product a specific result
- **Develop Use Cases**: document Use-Case diagrams, Activity Diagrams, etc
- **Develop Supplemental Specifications**: define requirements that cannot be readily captured by the use-case model such as system, regulatory requirements, application standards, etc. These specifications are captured in the Master Requirements document
- **Refine System Definition**: detail use cases and model and prototype the use interface (wire-frame, etc)

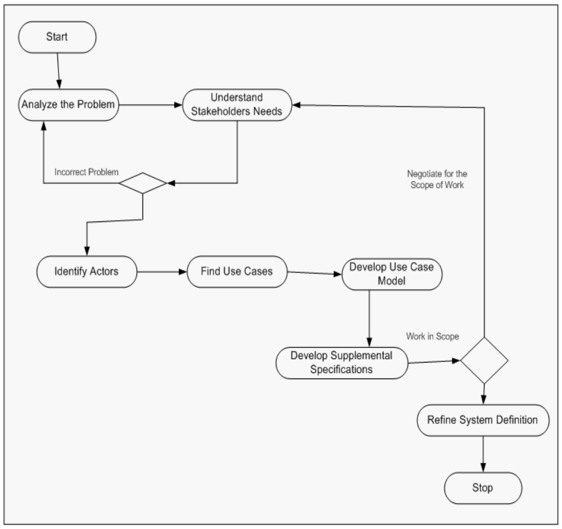


Figure 12: Requirements Flow

8. TESTING WORKFLOW

The following testing workflow was considered during the execution is depicted in the figure 13 and detailed test flow chart is depicted in figure 14.

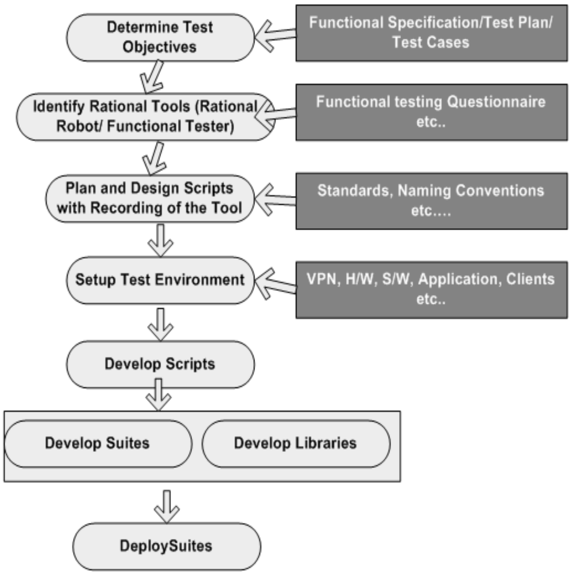

Determine Test Objectives
Functional Specification/Test Plan/ Test Cases
Identify Rational Tools (Rational Robot/ Functional Tester)
Functional testing Questionnaire etc..
Plan and Design Scripts with Recording of the Tool
Standards, Naming Conventions etc....
Setup Test Environment
VPN, H/W, S/W, Application, Clients etc..
Develop Scripts
Develop Suites
Develop Libraries
DeploySuites

Figure 13: Testing workflow

Key elements to be noted are:

- This is a generic test process flow for various types of tests to be executed during the project. Test method and type are determined after defining the test criteria.

- Testing can be executed on components, subsystems and the overall system, provided they are in a stable condition (stability)
- The test result will be tuned for another round of test cycle

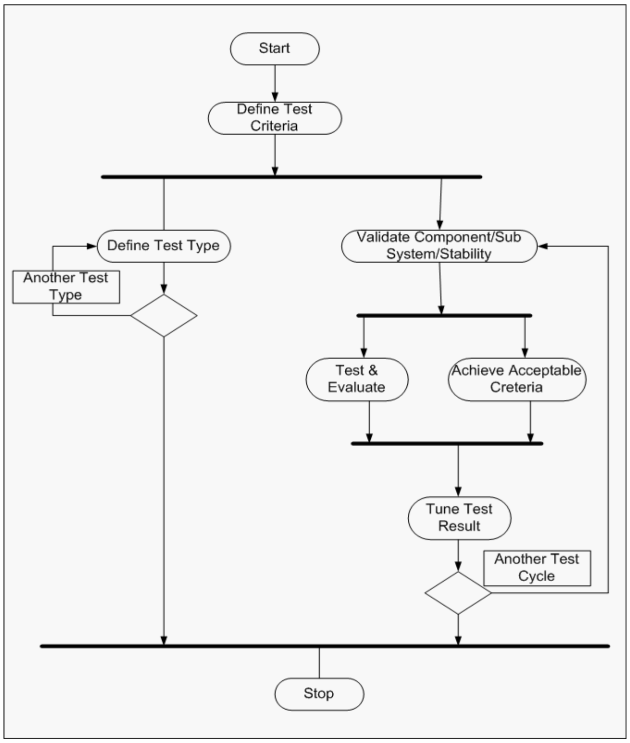


Figure 14: Test Detailed Flow chart

9. Analysis of Metrics

The detailed metric analysis has been performed on three different modules (EFT, REP, SD modules) over three cycles/builds with RUP implementation. The results obtained from the PoC, which is experimented for RUP implementation the number of defects are significantly reduced, in incremental cycles, which are analyzed in graphs depicted in figures 15 – 26 from data collected from the defect Consolidation log for three different modules EFT, REP and SD shown in table 9 - 11.

Defect Type: LG-Logic, CO-Computation, IF-Interface, UI-User interface, ST-Standards, CF-Configuration, DT-Data, SY-System, DC-Documentation, OT-Others

10. RESULTS AND REALIZED BENEFITS

Based on the results obtained from the prototype, it is evident that, downsize of the number of defects occurring in a module from iteration n to iteration (n+1). For example, the number of defects observed in the EFT module in the first iteration is reduced to in the second iteration and also the in the third iteration and also severity of defects. Also the reduction of the defects can also be seen in all the phases of the software development life cycle. The same trend was also observed for other two modules i.e. for REP and SD modules shown in tables 10 & 11 respectively along with the graphical representation. In this case study we have considered mainly "number of defects" parameter which is effecting the quality of service and hence the reliability.

Also with the current engagement model without RUP, the productivity is 120 LOC/PD. By using the customized RUP approach with test automation, the productivity is increased to 134 LOC/PD, which is approximately increased by 12%. Iterative development resulted in reducing the development duration, without waiting for the complete set of requirement. One set of requirements were developed in first iteration, while the second set of requirements were in elaborated phase, in parallel. By the time the first set finished its integrated testing, the development of second set is initiated. After doing an integration test on the second set, the complete requirements were tested in QAT. The duration of QAT effort was reduced drastically, from 3 months to 1 month. The number of defects in QAT also reduced by 90% compared to the PoC which was not executed based on RUP.

1. Iterative development enables process improvement and maturity in the first few iterations of the same project.
2. The PoC application integrates with another 20 applications which are not available at the time of PoC. Therefore, authors have created/used stubbed environment. This resulted frequent code drops to onsite for code builds into IST environment.
3. Risks are identified and mitigated well in advance. As mentioned in the above, offshore team has dependency not only on the external services, but also dependency on data. Authors recognized the dependency as high risk at the beginning. To mitigate the risk, a pre-IST period of two weeks was provided with the actual IST environment for testing the iterations. This facilitated in minimizing the errors during the IST testing cycle.

10.1 Recommendations

Based on the experience with several customer engagements, authors recommend the following to increase the operational efficiency and delivery excellence with special focus to defect density and other parameters indicated in the graphs.

10.2 Tool Driven Approach

Authors recommend using the following tools (from IBM) for delivery excellence.

Phase	**Tool**
Requirement Analysis	RequisitePro
Architecture	Rational Software Architect
Design	Rational Software Architect
Coding	Any IDE, should have facility for Rational Plug-in. For example Rational Application Developer **(RAD)**
Unit Testing	JUnit and NUnit
System Testing	Rational Functional

Phase	Tool
	Tester (RFT)
Performance Testing	Rational Performance Tester
Configuration Management	Rational ClearCase
Defect Management	Rational ClearQuest
RUP Implementation	RUP Builder / Rational Method Composer (RMC)
Documentation	Rational SoDA

Table 12: Recommended tools and technologies Effective Requirements Management

1. By capturing requirements with RequisitePro, the traceability can be made very simple as the same project can be imported into the design and testing.
2. During the development of the project, any changes need to done can be communicated to stakeholders by configuring mailing system with RequisitePro

10.3 Configuration Management

By having access to Clearcase from offshore, the following activities can be done efficiently.

10.3.1 Testing

Authors recommend early involvement in testing.

1. The test process should start as early as in the later stage of the inception phase, but after the first or second iteration of the requirements activities.
2. The majority of testing should be conducted during the construction phase. 10-20% (System, Load and Acceptance Test) will be executed during the transition phase.

10.3.2 Metrics and Measurement

Metrics are absolutely important to measure the project quality and performance. Authors recommend using Rational Test Manager to obtain various metrics to understand the project health. Test Manager also has an integrated reporting engine to generate various graphs and reports.

11. CONCLUSION

Based on the work that was carried by team and the obtained results, it is recommended to use RUP with careful customization to result in significant impact in productivity. In the inception phase of initial iterations, the productivity is low, but it will be improved in the subsequent iterations. In onsite-offshore model, it is recommended to have similar infrastructure to adopt the process efficiently with minimal/no dependency. This process also minimizes the risk as the dependent components would be addressed in the initial phases.

Based on the results obtained from PoC, which is experimented for RUP implementation, the number of defects is significantly reduced, in incremental cycles. This is due to the methodical approach suggested by RUP which is tool driven approach.

In mathematical modeling, the software reliability is inversely proportional to the number of defects. The results obtained in this paper are indicating the increase of reliability by reducing the number of defects.

REFERENCES

[1] P. Iyengar, Application Development Is More Global than Ever, publication G00124025,Gartner,2004; www.gartner.com/resources/124000/124025/application_dev.pdf.

[2] Hakan Erdogmus, National Research Council Canada, The Economic Impact of Learning and Flexibility on Process Decisions

[3] D.Sotirovski, "Heuristics for iterative software development", IEEE Software, May/June 2001, pp.66-73.

[4] Rational Unified Process: http://www.ibm.com

[5] P. Kruchten, the Rational Unified Process: An Introduction, 2e. Addison-Wesley-Longman, 2000

[6] Hans Westerheim, Geir Kjetil Hanssen: The Introduction and Use of a Tailored Unified Process: 2005 IEEE Software

[7] Ivar Jacobson, Shaping Software Development: May/June 2002 IEEE Software

[8] J. G. Proakis and D. G. Manolakis – *Digital Signal Processing – Principles, Algorithms and Applications*; Third Edition; Prentice Hall of India, 2003.

Cycle 1																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	EFT	2	10	5	140Pg	0	1	1	1								1		2				
Design	EFT	5	15	25	100Pg	5	0	0	5								1		1	4			
Coding	EFT	18	10	20	200Kloc	1	11	6	2				16							3	15		
Unit Testing	EFT	75	75	150	200Kloc	20	20	35	25	10	5	15	15		5							75	
IST Support	EFT	11	10	60	20Kloc	1	3	7	2			5			1	3							11
Cycle 2																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	EFT	1	5	6	160Pg	0	0	1	1										1				
Design	EFT	2	10	20	130Pg	2	0	0	2								1		1	1			
Coding	EFT	9	15	30	260Kloc	0	6	3	0				9							1	8		
Unit Testing	EFT	25	16	40	258Kloc	5	9	11	5	4	4	7	4		1							25	
IST Support	EFT	7	10	20	20Kloc	0	1	6	1			2			1	3							7
Cycle 3																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	EFT	0	3	4	190 Pg	0	0	0	0										0				
Design	EFT	0	15	25	180 Pg	0	0	0	0								0		0	0			
Coding	EFT	2	10	48	325 Kloc	0	1	1	0				2							1	1		
Unit Testing	EFT	5	75	42	200Kloc	0	2	3	0	2	0	2	1		0							5	
IST Support	EFT	1	10	30	20Kloc	0	0	1	0			0			0	1							1

Table 9: Defect consolidation log for EFT Module

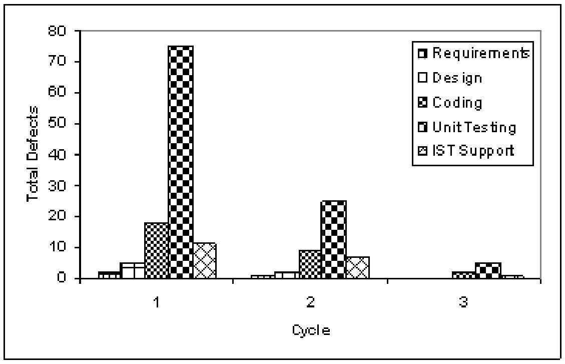


Figure 15: Total defects analysis for EFT module

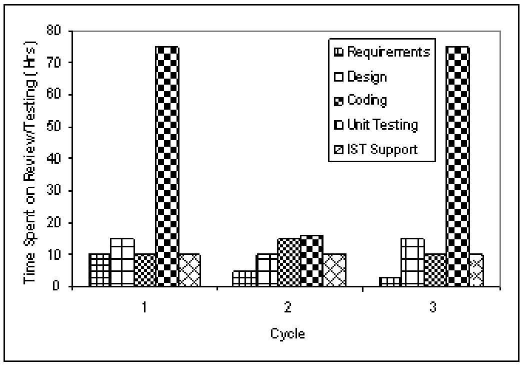


Figure 16: Time spent on Review/testing (Hrs). analysis for EFT modul

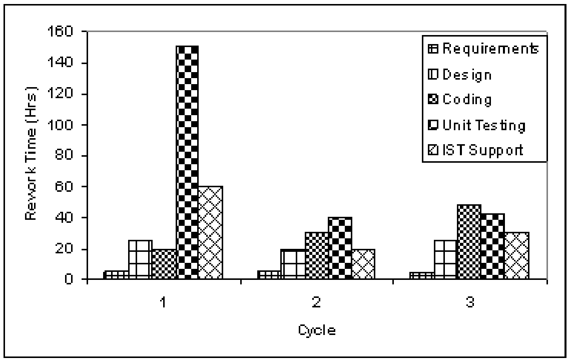


Figure 17: Rework Time (Hrs) analysis for EFT module

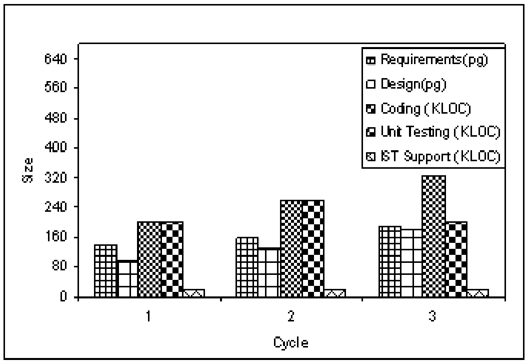


Figure 18 Size (pg/KLOCS) analysis for EFT module

Cycle 1																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	REP	2	5	5	150Pg	0	2		1								1		2				
Design	REP	3	5	5	36Pg	0	3	0	3										1	2			
Coding	REP	6	10	10	100Kloc	2	2	2	2				4							1	5		
Unit Testing	REP	20	40	80	100Kloc	5	10	5	5	5		10			5							20	
IST Support	REP	0	10	0	6Kloc	0	0	0															
Cycle 2																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	REP	1	2	9	163 Pg	0	1										1		1				
Design	REP	1	5	5	45 Pg	0	1	0	1										0	1			
Coding	REP	2	8	21	138 Kloc	0	1	1	1				1							0	2		
Unit Testing	REP	9	10	14	100Kloc	2	5	2	1	2		4			2							9	
IST Support	REP	0	10	0	6Kloc	0	0	0															
Cycle 3																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	REP	0	1	5	192 Pg	0	0										0		0				
Design	REP	0	5	1	49Pg	0	0	0	1										0	0			
Coding	REP	1	14	10	170 Kloc	0	0	1	1											0	1		
Unit Testing	REP	1	10	13	110 K Loc	0	0	1	1	0		0			0							1	
IST Support	REP	0	10	0	6Kloc	0	0	0															

Table 10: Defect consolidation log for REP Module

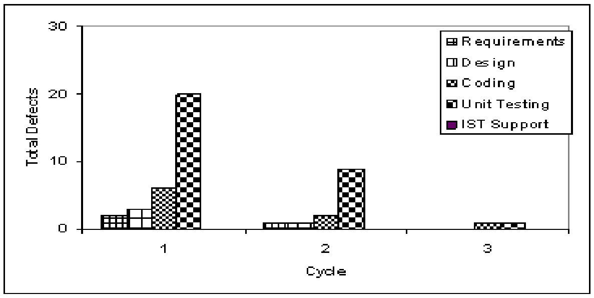


Figure 19: Total defects analysis for REP module

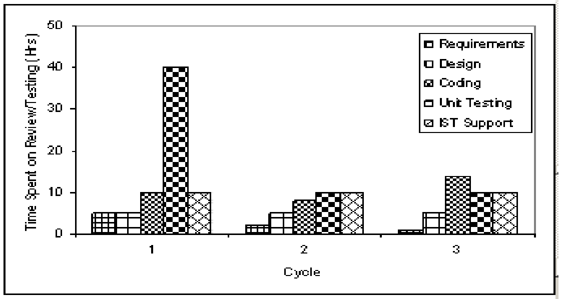


Figure 20: Time spent on Review/testing (Hrs) analysis for REP module

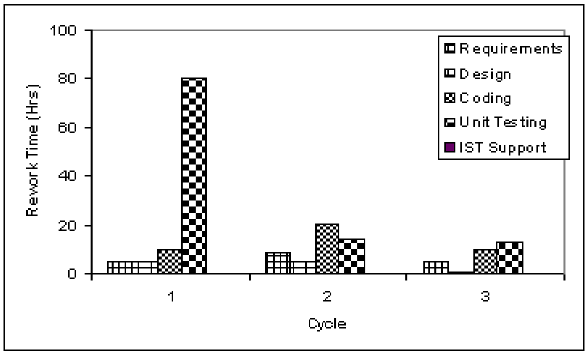


Figure 21: Rework Time (Hrs) analysis for REP module

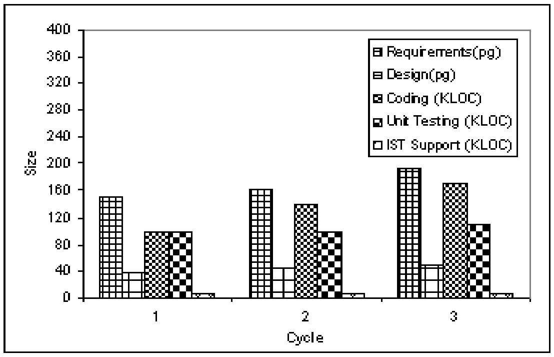


Figure 22: Size (pg/KLOCS) analysis for REP module

Cycle 1																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	SD	3	15	10	300Pg	0	1	2	3										3				
Design	SD	3	15	15	70Pg	0	1	2	3										1	2			
Coding	SD	20	10	30	150Kloc	4	6	10	4				15				1			4	16		
Unit Testing	SD	45	40	70	150Kloc	10	15	20	10	15		20			5							45	
IST Support	SD	4	15	30	10Kloc	0	1	3				3			1								4
Cycle 2																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin.				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	SD	2	10	8	315 pg	0	1	1	2										2				
Design	SD	2	11	10	70Pg	0	1	1	2										0	2			
Coding	SD	11	8	16	190 Kloc	1	4	6	2				9				0			1	10		
Unit Testing	SD	14	17	30	156 KLOC	1	9	4	2	5		6			1							14	
IST Support	SD	1	5	15	10Kloc	0	0	1				1			0								1
Cycle 3																							
Phase	Unit / Module	Total Defects	Time Spent on Review / Testing Hrs	Re-work Time (Hrs)	Size	No. of Defects - by Severity			No of Defect – by Type										No. of Defects – by Phase of Origin .				
						H	M	L	LG	CO	IF	UI	ST	CF	DT	SY	DC	OT	Require ments	Desi gn	Coding	Testing	IST Support
Requirements	SD	0	5	3	290 Pg	0	0	0	0										0				
Design	SD	0	9	9	70Pg	0	0	0	0										0	0			
Coding	SD	2	8	14	205 KLOC	0	1	1	1				1				0			0	2		
Unit Testing	SD	2	14	26	163 KLOC	0	1	1	0	0		1			1							2	
IST Support	SD	0	5	15	10Kloc	0	0	0				0			0								0

Table 11: Defect consolidation log for SD Module

Continued on page no. 60

NFCKE: New Framework for Document Classification and Knowledge Extraction

Ghanshyam Singh Thakur[1] and **Dr. R. C. Jain**[2]

Abstract - In this research paper we have developed NFCKE text categorization systems. Text classification has many applications, such as fraud detection, automatic email classification and web-page categorization. The results show that NFCKE does better than other widely used techniques. The new framework for document classification is efficient and accurate. Our experiments indicate that the accuracy of existing method increase by using this approach. The final goal is achieving high performance and eventually increasing classification accuracy.

Index Terms - Text mining, classification, Binary Matrix Model (BMM).

1. INTRODUCTION

Document classification has been studied intensively because of its wide applicability in areas such as web mining, information retrieval. The majority of this information is in text format, for example, emails, news, web pages, reports, etc. Organizing them into a logical structure is a challenging task. More recently, classification is employed for browsing a collection of documents or organizing the query results. Although standard classification techniques such as k-means, Support Vector Machines, Naive Bayes, Decision Trees[15,16,18,19], can be applied to document classification, they usually do not satisfy the special requirements for classification documents: all these approaches are suffer from lack of high performance and high accuracy. In addition, many existing document classification algorithms require the user to specify the number of category as an input parameter. Incorrect estimation of the value always leads to poor classification accuracy. Furthermore, many classification algorithms are not robust enough to handle different types of document sets in a real-world environment. In some document sets, category sizes may vary from few to thousands of documents. This variation tremendously reduces the resulting classification accuracy for some of the state-of-the art algorithms. But there are still problems to be tackled such as efficiency and accuracy. Owing to wide significant applicability of text categorization and challenges in the area motivated us to do work in this field. The poor classification accuracy and the weaknesses of the standard classification methods formulate the goal of this research. We provide an accurate, efficient, and scalable classification method that addresses the special challenges of document classification. NFCKE is a relatively new concept comparatively more efficient and accurate. So for, no research was conducted to use NFCKE concept for Document classification. This approach seems promising because we can apply different method of classification. BMM is the out come of NFCKE. We applied existing classification methods on the BMM. Our experiments indicate that the accuracy of existing method increase by using the model (BMM). We can also apply associative classifier on BMM Model, which generate association rules between keywords or features. The final goal is achieving high performance and eventually increasing classification accuracy. All these issues motivated and directed our research.

2. METHODOLOGY

The main goal of this research is to develop high performance and new optimization text categorization algorithms that will reduce the time complexity and space complexity of algorithms and finding various applications of the text categorization algorithm in real world problem as the result of computerization. The amount of text documents available in digital form has been growing significantly during the last decades due to the development of new technology. These include e-mail, newsgroups and on-line news, all of which can be stored in text form. The accelerating growth in the amount of text data makes it necessary to automate. In this paper we limit our attention to document classification accuracy and high performance.

Text classification is based on supervised learning model. In this learning we divided our dataset into two parts. One part is called training dataset and another part is called test dataset. With training dataset we create a model or classifier. Once we created a classifier we estimate the accuracy of the classifier using test dataset. For mining large document collections it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file. The aim of this paper is to preprocess documents, to apply classification method and improve accuracy of text categorization using NFCKE, which is the basis for various applications such as e-mail classification and Web-page classification. NFCKE is a new concept for Text categorization witch includes several sub-phases that should be integrated for efficient and accurate outcomes. These sub phases include document collection, preprocessing, indexing, feature selection, model preparation and estimation of classifier accuracy and performance. We explore these sub-phases in terms of an approach to similarity based text categorization. After performing sub phases like preprocessing, indexing and feature selection of NFCKE, we apply BMM-based text classification method because, which is fast and more robust method compared to others text classification methods. BMM-based text classification is one of the new supervised approaches to classify texts into a set of predefined classes with relatively low computation. New framework documents are almost unstructured and all classifications algorithm requires structured form of the

[1, 2] *Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha(M.P.), INDIA*

documents. So it is the mandatory to convert unstructured document into a structured document representation. In our new framework we explain in details this conversion. This new framework includes following sub-phases-

1. Document Collection-we collect relevant documents for classification.
2. Preprocessing-in this phase we perform the operation like removal of HTML Tags, Special character, stop words and perform word stemming
3. Indexing- in this phase we perform the operation of assigning the weight to feature
4. Feature Selection- in this phase we perform reduce dimension of documents
5. BMM Representation- to represent classes into two dimension tabular form.
6. Performance and accuracy: we estimate the performance and accuracy of BMM method

After performing sub phases like preprocessing, indexing and feature selection of NFCKE, we apply BMM-based text classification method because, which is fast and more robust method compared to others text classification methods. BMM-based text classification is one of the new supervised approaches to classify texts into a set of predefined classes with relatively low computation

3. DOCUMENT REPRESENTATION

All the algorithms applied in Text Classification need the documents to be represented in a way suitable for the inducing of the classifier. For the document classification we represent using Binary Matrix Model (BMM) as binary matrix. In Binary Matrix each rows represent a documents and each columns represent term in the documents. Number of rows indicates number of document and number of column indicate number of terms in the document.

3.1 Binary Matrix Model (BMM)-

This model represent by a matrix called binary matrix.Binary Matrix M is represented as

➢ M[di × wj] = 1 , if wj Є di
➢ = 0, otherwise
➢ Where i=1,2,…..n
➢ j=1,2,3,…..m

Binary Matrix Model (BMM) is a powerful approach. This model is based on the binary values, 0 represents the absence of the term in the document and 1 represents presence of the term in the document. This model can used wide variety document classification algorithm.

4. EXPERIMENTAL RESULT ANALYSIS

We have done the experiment on 20 Newsgroups data sets— This dataset is a collection of 20,000 newsgroup documents, partitioned into 20 different newsgroups. The 20 different news groups are as follows -
alt.atheism,comp.graphics,comp.os.ms-windows.misc, talk.politics.misc, comp.sys.mac.hardware,comp.windows.x,misc.forsale,rec.autos,rec.motorcycles,rec.sport.baseball,rec.sport.hockey,sci.crypt,sci.electronics,sci.med,sci.space,soc.religion.christian,talk.politics.guns,talk.politics.mideast,talk.religion.misc,comp.sys.ibm.pc.hardware

S.No.	BMM	k-nn	Bayes
Datase1	91.81	83.98	84.98
Datase2	92.16	84.14	86.14
Datase3	91.17	84.76	85.76

Table 1

To improve the classification performance, we adopt the following method each data set was split into 80% and 20% for training and testing respectively. A 3-fold cross validation was carried out to determine the training set accuracy at the various parameter settings. The parameter that gave the best training accuracies was then used to determine the accuracy values for each of the corresponding test sets. To compare the performance of the classification methods, we look at a set of standard performance measures. .

Figure 1

In this learning we divided our dataset into two parts. One part is called training dataset and another part is called test dataset. With training dataset we create a model or classifier. Once we created a classifier we estimate the accuracy of the classifier using test dataset. BMM-based classifier consistently and substantially outperforms other algorithms such as Naive Bayesian, k-nearest-neighbors, and C4.5, on a wide range of datasets. BMM -based text classification presents accuracy close to that of the state-of-the-art methods.

5. CONCLUSION

In this paper we have developed a new framework for classification and a new BMM classifier. We conducted extensive comparative experiments on standard test collections (the 20-Newsgroups). We experimentally predict that a BMM model which is the outcome of NFCKE give high accuracy and efficiency for classification. We also experimentally showed that BMM gives high accuracy and performance than other classification methods. The results show that NFCKE does better than other widely used techniques.

Continued on page no. 59

Data Dissemination in Mobile Computing Environment

S Krishna Mohan Rao[1] and **Dr. A Venugopal Reddy**[2]

Abstract - ***Data dissemination in asymmetrical communication environment, where the downlink communication capacity is much greater than the uplink communication capacity, is best suited for mobile environment. In this architecture there will be a stationary server continuously broadcasting different data items over the air. The mobile clients continuously listen to the channel and access the data of their interest whenever it appears on the channel and download the same. The typical applications of such architecture are stock market information, weather information, traffic information etc. The important issue that is to be addressed in this type of data dissemination is – how quickly the mobile clients access the data item of their interest i.e. minimum access time so that the mobile clients save the precious battery power while they are on mobile. This paper reviews the various techniques for achieving the minimum access time. The advantages and disadvantages are discussed and explored different research areas for achieving the minimum access time***.

1. INTRODUCTION

There are two fundamental information delivery methods for wireless data applications: Point-to-Point access and Broadcast. Compared with Point-to-Point access, broadcast is a more attractive method. A single broadcast of a data item can satisfy all the outstanding requests for that item simultaneously. As such, broadcast can scale up to an arbitrary number of users. There are three kinds of broadcast models, namely *push-based* broadcast, *On-demand* (or *pull-based*) broadcast, and *hybrid* broadcast. In push based broadcast [1, 2], the server disseminates information using a periodic/aperiodic broadcast program (generally without any intervention of clients). In on demand broadcast [3, 4], the server disseminates information based on the outstanding requests submitted by clients;

In hybrid broadcast [5, 6, 7], push based broadcast and on demand data deliveries are combined to complement each other. Consequently, there are three kinds of data scheduling methods (i.e., *push based scheduling*, *on demand scheduling*, and *hybrid scheduling*) corresponding to these three data broadcast models.

2. PUSH BASED DATA SCHEDULING

In push based data broadcast, the server broadcasts data proactively to all clients according to the broadcast program generated by the data scheduling algorithm. The broadcast program essentially determines the order and frequencies that the data items are broadcast in. The scheduling algorithm may make use of precompiled access profiles in determining the broadcast program. In the following, two typical methods for push based data scheduling are described, namely *flat broadcast* and *broadcast disks.*

Flat Broadcast The simplest scheme for data scheduling is flat broadcast. With a flat broadcast program, all data items are broadcast in a round robin manner. The access time for every data item is the same, i.e., half of the broadcast cycle. This scheme is simple, but its performance is poor in terms of average access time when data access probabilities are skewed.

Broadcast Disks Hierarchical dissemination architecture, called *Broadcast Disk* (Bdisk), was introduced in [1]. Data items are assigned to different *logical* disks so that data items in the same range of access probabilities are grouped on the same disk. Data items are then selected from the disks for broadcast according to the relative broadcast frequencies assigned to the disks. This is achieved by further dividing each disk into smaller, equal size units called *chunks*, broadcasting a chunk from each disk each time, and cycling through all the chunks sequentially over all the disks. A *minor cycle* is defined as a sub cycle consisting of one chunk from each disk. Consequently, data items in a minor cycle are repeated only once. The number of minor cycles in a broadcast cycle equals the Least Common Multiple (LCM) of the relative broadcast frequencies of the disks. Conceptually, the disks can be conceived as real physical disks spinning at different speeds, with the faster disks placing more instances of their data items on the broadcast channel.

However, if the number of minor cycles in a broadcast cycle is not equal the Least Common Multiple (LCM) of the relative broadcast frequencies of the disks, dividing precisely the desired number of chunks, is a problem. [13] addressed this problem by suggesting to fill up the disk with other information and making it divisible so that the number of minor cycles is equal to the LCM of relative broadcast frequencies.

3. ON-DEMAND DATA SCHEDULING

Push based wireless data broadcasts are not tailored to a particular user's needs but rather satisfy the needs of the majority. Further, push-based broadcasts are not scalable to a large database size and react slowly to workload changes. To alleviate these problems, many recent research studies on wireless data dissemination have proposed using on-demand data broadcast (e.g., [3, 4, 8, 9]). A wireless on demand broadcast system supports both broadcast and on demand services through a broadcast channel and a low bandwidth uplink channel. The uplink channel can be a wired or a wireless link. When a client needs a data item, it sends to the server an on demand request for the item through the uplink. Client requests are queued up (if necessary) at the server upon arrival. The server repeatedly chooses an item from among the

[1]Head-MCA, Mahaveer Institute of Science & Technology, Bandlaguda, (Post): Kesavagir, Hyderabad – 500 005
[2]Professor & Dean – CSE, O U College of Engineering, Osmania University Hyderabad – 500 007
E-Mail: [1]skrishnamohanrao@yahoo.com

outstanding requests, broadcasts it over the broadcast channel, and removes the associated request(s) from the queue. The clients monitor the broadcast channel and retrieve the item(s) they require.

The data-scheduling algorithm in on demand broadcast determines which request to service from its queue of waiting requests at every broadcast instance.

4. HYBRID DATA SCHEDULING

Push-based data broadcast cannot adapt well to a large database and a dynamic environment. On-demand data broadcast can overcome these problems. However, it has two main disadvantages: i) more uplink messages are issued by mobile clients, thereby adding demand on the scarce uplink bandwidth and consuming more battery power on mobile clients; ii) if the uplink channel is congested, the access latency will become extremely high. A promising approach, called hybrid broadcast, is to combine push-based and on-demand techniques so that they can complement each other. In the design of a hybrid system, one of the main issues is the assignment of a data item to push-based broadcast, on-demand broadcast or both.

Concerning this issue, there are different proposals for hybrid broadcast in the literature. In the following, we introduce the techniques for balancing push and pull and adaptive hybrid broadcast.

Balancing Push and Pull: Hybrid architecture was first investigated in [10, 11]. In that model, items are classified as either frequently requested (*frequest*) or infrequently requested (*irequest*). It is assumed that clients know which items are *frequests* and which are *irequests*.

The model services *frequests* using a broadcast cycle, and *irequests* using on-demand. In the downlink scheduling, the server makes consecutive transmissions of frequented items (according to a broadcast program), followed by the transmission of the first item in the *irequest* queue (if at least one such request is waiting). Analytical results for the average access time were derived in [11].

In [5], the push based *Bdisk* model was extended to integrate with a pull based approach. The proposed hybrid solution, called *Interleaved Push and Pull* (*IPP*), consists of an uplink for clients to send to the server pull requests for the items that are not on the push-based broadcast. The server interleaves the *Bdisk* broadcast with the responses to pull requests on the broadcast channel.

The disadvantage of this approach is that if there is not enough bandwidth for pulls, the performance might degrade severely since the pull latencies for non-broadcast items will be extremely high.

In [14], an attempt was made to compare various broadcast scheduling algorithms. For this, a simulation model, Sketch-it, is developed and compared various algorithms. This is very useful for conducting various experiments by changing the critical parameters.

In [15], the multicast server offering the data items at a variety of transmission speeds to the clients' varied requests is discussed. The paper proposes to slice a server's available outgoing network capacity in to data channels, assign server's data to those channels, and assign clients to the channels given client' varied requests and download speeds.

5. DATA ALLOCATION OVER MULTIPLE BROADCAST CHANNELS

Multiple physical channels have capabilities and applications that can not be mapped on to single channels. As stated in [12] some example advantages include better fault tolerance, configurability and scalability. By having access to multiple physical channels fault tolerance is improved. For example if a server broadcasting on a certain frequency crashes, its work must be migrated to another server. If this server is already broadcasting on another frequency it can only accept the additional work if it has the ability to access multiple channels. More flexibility is allowed in configuring broadcast servers. Assume that there are two contiguous cells, which contain broadcast servers that transmit at different channels. A single server that wishes to take over the responsibility of transmitting in both cells can only do so if it can transmit over multiple channels. Finally, being able to transmit over multiple channels has scalability benefits. A broadcasting system must be able to handle both high powered and low powered clients. In order to do so, multiple channels can be used and clients can monitor a number of channels commensurate to their capacities and data needs.

This calls for a data-scheduling algorithm, which works dynamically, and allocates data according to changing access patterns to achieve efficient data access and channel utilization so that the access time is minimum. However, the area of interest is hoe to adjust the broadcast program when the data items are changing dynamically. This calls for a research on incremental algorithms to change the programs dynamically.

[16] explores the problem of adjusting broadcasting programs to effectively respond to the changes of data access frequencies, and develop an efficient algorithm DL to address this aspect. The DL algorithm showed the high quality of results and close to the optimal ones.

[17] explains the effects of dynamicity on broadcast program with respect to *item placement, Disk structure, Disk content* and *Disk Values. Item placement or Disk Structure* changes the relative frequencies and/ or order of appearance of data items already being broadcast. The *value of Data item* changes only when it is updated. Dynamicity due to *Disk contents* does not influence the items that appear on broadcast.

[18] explores the problem of dynamic data and channel allocation with the number of communication channels and the number of data items are given. Algorithm SOM is a composite algorithm which will cooperate with 1) a search strategy and 2) a broadcast program generation. However the algorithm is not easy for implementation.

[10] proposes optimal allocation algorithm which searches exhaustively to find the optimal solution for channel allocation and data page organisation. However, the execution of algorithm is very slow

6. CONCLUSION & FUTURE STUDY

This paper discusses various techniques for data dissemination in mobile communication environments. Data scheduling methods are investigated with respect to their performance in minimum access time. For data scheduling push based, on-demand, hybrid, and multi channel broadcast were discussed. Push based broadcast is attractive when access patterns are known before hand, while on-demand broadcast is desirable for dynamic access patterns. Hybrid data broadcast offers more flexibility by combining push-based and on-demand broadcasts. Broadcasting in multi channel does have advantages in terms of high fault tolerance. The research areas like scheduling data items dynamically by employing incremental algorithms are identified. Another research area of interest is how the server gets the feedback from the mobile clients regarding their access patterns so that it adjusts the scheduling program accordingly. The authors are working on developing incremental algorithm for the multi channel data scheduling so that the server program adjusts itself to the dynamically changing data access patterns of the mobile clients.

REFERENCES

[1] S. Acharya, R. Alonso, M. Franklin, and S. Zdonik. Broadcast disks: Data management for asymmetric communications environments. In Proceedings of ACM SIGMOD Conference on Management of Data, pages 199–210, San Jose,CA, USA, May 1995.

[2] S. Hameed and N. H. Vaidya. Efficient algorithms for scheduling data broadcast. ACM/Baltzer Journal of Wireless Networks (WINET), 5(3):183–193, 1999.

[3] S. Acharya and S. Muthukrishnan. Scheduling ondemand broadcasts: New metrics and algorithms. October 1998.

[4] D. Aksoy and M. Franklin. R x W: A scheduling approach for largescale ondemand data broadcast. IEEE/ACM Transactions on Networking, 7(6):846–860, December 1999.

[5] S. Acharya, M. Franklin, and S. Zdonik. Balancing push and pull for data broadcast. In Proceedings of ACM SIGMOD Conference on Management of Data, pages 183–194, Tucson, AZ, USA, May 1997.

[6] T. Limielinski and S Viswanathan. Adaptive wireless information systems. In proceedings of the Special Interest Group in Database Systems (SIGDBS) Conference, Pages 19-41, Tokyo, Japan, October 1994

[7] W.-C.Lee, Q. L. Hu, and D. L. Lee. A study of channel allocation methods for data dissemination in mobile computing environments. ACM/Baltzer Journal of Mobile Networks and Applications (MONET), 4(2):117–129, 1999.

[8] Q L Hue, D L Lee and W C Lee. Performance evaluation of a wireless hierarchial data dissemination system. Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom'99), pages 163-173, Seatle, WA,USA, August 1999.

[9] C J Su, L. Tassiulas and V J Tsotras. Broadcast scheduling for information distribution. ACM/Baltzer Journal of Wireless Networks (WINET), 5(2):137-147, 1999

[10] J W Wong, Broadcast delivery. Proceedings of the IEEE,76(12):1566-1577, December 1998

[11] J W Wong and H D Dykeman, Architechtecture and Performance of large scale information delivery networks. In Proceedings of the 12th International Teletraffic Congress, pages 440-446, Torino, Italy, June 1988.

[12] Wai Gen Yee, Shanmukhanath, B Navathe, Edward Omaian Ciniski and Christopher Jermaine. Efficient Data Allocation over multiple channel at Broadcast Servers. IEEE Transactions on Computers, Volume 51, No 10, October 2002.

[13] Karl Aberer, Data Broadcasting in Mobile Networks, EPFL-SSC, laboratore de systemes d'informations repartis, pages 1-38, 2004

[14] Alexander Hall and Hanjo Taubig, Comparing Push- and Pull- Based Broadcasting. www14.in.tum.de/publicationen/2003

[15] Wang Lam and Hector Garcia-Molina, Slicing Broadcast Disks, 2003

[16] Een-Chih Peng, Jiun-Long Huang and Ming Syan Chen, Dynamic Levelling: Adaptive Data Broadcasting in a Mobile Computing Environment. Mobile Networks and Applications 8,355-364, 2003

[17] R K Ghosh, Data Dissemination, Mobile Computing, IIT – Kanpur, pages 1-70, 2006

[18] jiun-Long Huang, Wen-Chih Peng and Ming-Syan Chen. SOM: Dynamic Push-Pull Channel Allocation Framework for Mobile Data Broadcasting. IEEE Transactions on Mobile Computing Vol 5, No. 8 August 2006

[19] S Krishna Mohan Rao and Dr A Venugopal Reddy: Optimal allocation algorithm for data broadcasting programs in a mobile computing environment. Proceedings of National Conference INDIACom-2007.

Continued from page no. 56

REFERENCES

[1] A.K. Pujari 2002. "Data Mining Techniques", University press 2002.

[2] Ana Cardoso-Cachopo Arlindo L. Oliveira 2006, "Empirical Evaluation of Centroid-based Models for Single-label Text Categorization",INESC-ID Technical Report 2006

[3] Ken Williams 2006, "A Framework for Text Categorization", Web Engineering Group The University of Sydney Bldg J03, Sydney NSW 2006

[4] Makoto 2006, "Hierarchical Bayesian Clustering for Automatic Text Classification", Proceedings of the Fifth

International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006

[5] Andreas Hotho 2005. "A Brief Survey of Text Mining" KDE Group University of Kassel May 13, 2005

[6] Fafal Rak, Wojciech Stach 2005."Considering Re-occurring Features in Associative Classifiers", Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005. Proceedings

[7] Goutte, C. and Gaussier, E. 2005, "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation." In Pro-ceedings of the 27th European Conference on Information Retrieval, pages 345-359.

[8] Huan Liu, and Lei Yu 2005. "Toward Integrating Feature Selection Algorithms for Classification and Clustering", Knowledge and Data Engineering, IEEE Transactions on April 2005 Vol. 17, pages 491- 502

[9] Huang, J. and Ling, C. 2005, "Using AUC and Accuracy in Evaluating Learning Algorithms". IEEE Trans. on Data and Knowledge Engineering,17 3 3 :299-310.

[10] Kathrin Eichler 2005, "Automatic Classification of Swedish Email Messages",17th August 2005

[11] Kiritchenko, S., Matwin, S., Nock, R.,and Famili, A. 2005, "Learning and Evaluation in the Presence of Class Hierarchies": Application to Text Categorization. Submitted.

[12] Mohammed J. Zaki 2005, "Efficient Algorithms for Mining Closed Item sets and Their Lattice Structure" IEEE transactions on knowledge and data engineering, vol. 17, no. 4, april 2005

Continued from page no. 54

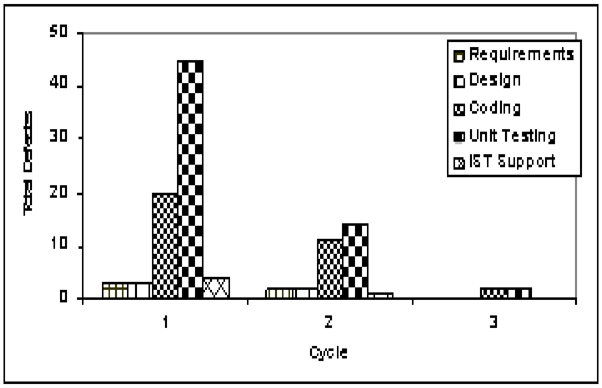


Figure 23: Total defects analysis for SD module

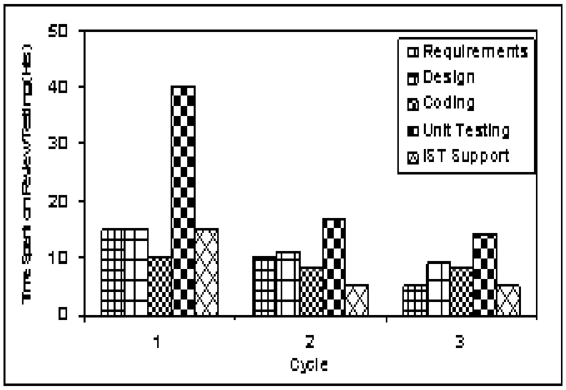


Figure24: Time spent on Review/testing (Hrs) analysis for SD module

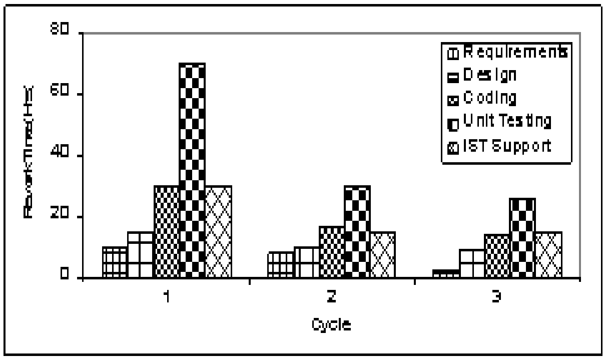


Figure 25: Rework Time (Hrs) analysis for SD module

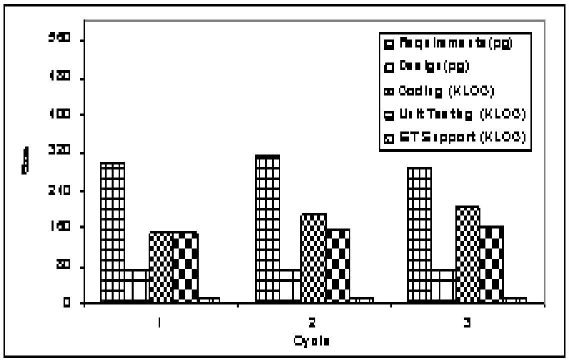


Figure 26: Size (pg/KLOCS) analysis for SD module

Design of an Agent Based Context Driven Focused Crawler

Naresh Chauhan[1] and **A.K. Sharma**[2]

Abstract - A focused crawler downloads web pages that are relevant to a user specified topic. Most of the existing focused crawlers are keyword driven and do not take into account the context associated with the keywords. This leads to retrieval of a large number of web pages irrespective of the fact whether they are logically related. Thus, the keyword based strategy alone is not sufficient for the design of a focused crawler as context relevance is more important as far as the user's requirement is concerned. This paper proposes the design of a context driven focused crawler (CDFC) that searches and downloads only highly related web pages, thereby reducing the network traffic. It also employs a category tree which is a flexible user interface showing the broad categories of the topics on the web. Since CDFC downloads only the relevant and credible documents, a very small number in comparison, the proposed design significantly reduces the storage space at the search engine side.

Index Terms - Search engine, Crawler, Hypertext Document System, Category Tree, Software Agents

1. INTRODUCTION

The World Wide Web (WWW) is a continuously expanding large collection of hypertext documents [1]. It represents a very large distributed hypertext system, involving hundreds of thousands of individual sites. It is a client-server based architecture that allows a user to initiate search by providing keywords to a search engine, which in turn collects and returns the required web pages from the Internet. Due to extremely large number of pages present on the web, the search engine depends upon crawlers for the collection of required pages. A Crawler [2] follows hyperlinks present in the documents to download and store web pages for the search engine.

Current commercial search engines maintain large number of web pages [3,7] and easily find several thousands of matches for an average query. Therefore, a search engine may present a list of thousands of web pages in response to user's particular keyword possibly consisting of irrelevant web pages also. The web search engines try to cover the whole web and serve queries concerning all possible topics [4]. In fact, from the user's point of view, it does not matter whether the search returned 10,000 or 50,000 hits because the number of matches becomes too large to sift, leading to the problem of *information overkill.*

The search quality of web pages can be improved by focused crawling [5,6,12] which aim to search and retrieve only the subset of the WWW that pertains to a specific topic of relevance. Focused crawler, therefore, offers a potential solution to the problem of information overkill. The existing focused crawlers [6,7] adopt different strategies for computing the words' frequency in the web documents. If higher frequency words match with the topic keyword, then the document is considered to be relevant. But the current crawlers are not able to analyze the context of the keyword in the web page before they download it. For instance, the word 'spider' has various interpretations. To a web programmer, it is the name of a software program used in search engines; to a general computer user it denotes a game of cards and to a layman it is simply name of an insect. Thus, the topical relevance is not the only issue for focused crawlers but context relevance should also be considered [10]. If the user issues one keyword then its relevant context must also be known.

In this paper, the design of a *Context Driven Focused Crawler* (CDFC) is being proposed that provides the context of the keywords to the user in a flexible and interactive category tree [5]. The agent-based design not only overcomes the complex time-consuming computations of existing focused crawlers but also reduces network traffic significantly. In line with the demands of a focused crawler that the relevant information should be collected and retrieved by the user in the least amount of time possible, the proposed architecture reduces the search time for a document and the information database on the search engine side becomes more easily manageable.

2. RELATED WORK

A similarity based crawler that orders URLs having target keyword in anchor text or URL, was probably one of the first efforts towards focused crawling [9]. The basic focus was to crawl more important pages first i.e. to look at various measures of importance for a page such as similarity to a driving query, number of pages pointing to this page (back links), page rank, location, etc. The Page Rank algorithm [11]
computes a page's score by weighing each in-link to the page proportionally to the quality of the page containing the in-link. Thus, a web page will have a high page rank, if the page is linked from many other pages, and the scores will be even higher if these referring pages are also good pages, i.e. having high Page Rank scores. In the HITS (Hyper-link-induced- topic search) algorithm [8], an authority page is defined as a high quality page related to a particular topic or search query and a hub page is one that provides pointers to other authority pages. Based upon this, a web page is associated with an *Authority Score* and a Hub Score that is calculated to identify the web page context.

Another focused crawler [7] employs seed keywords which are used to find seed URLs from some standard search engine like

[1]Asst. Prof., Deptt. of Computer Engg., YMCA Institute of Engineering, Faridabad - 121006, India
[2] Professor & Head, Deptt. of Computer Engg., YMCA Institute of Engineering, Faridabad - 121006, India
E-Mail: [1]nareshchauhan19@yahoo.com and 2ashokkale2@ rediffmail.com

Google. The seed URLs are used to fetch seed pages with the help of TF.IDF algorithm, based on iteratively calculating word frequency. This algorithm is used to find out some more number of keywords from seed web pages to represent the topic. Afterwards, vector similarity is computed between web page and topic keywords to see whether the page is relevant to the topic.

Diligenti et al [6] uses a general search engine to get the web pages linking to a specific document and builds up a context graph for the page. The graph is then used to train a set of classifiers to assign documents to different categories based on their expected link distance to the target. In fact, graphs and classifiers are constructed for each seed document with layers being built up to a specified level. Thus, the crawler gains knowledge about topics that are directly or indirectly related to the target topic.

A critical look at the available focused crawlers [5-9,11] indicates that these crawlers suffer from the following drawbacks :

I. The problem of iterative computation of word frequency for every web document renders the search process expensive.
II. The relevance of web page is not known until it is downloaded.
III. Associated context of the web page is unknown prior to search initiation.
IV. The user interface of a search engine with keyword search is not flexible.

The proposed work paper effectively addresses the above-mentioned issues. A Context driven focused crawler has been designed, which uses augmented hypertext document structure coupled with a category tree for providing user interface at the search engine side.

1.1 Augmented Hypertext Documents

The information on WWW is organized in the form of a large, distributed and non-linear text system known as Hypertext Document system. HTTP and HTML provide a standard way of retrieving and presenting the hyper-linked documents. The XML offers more flexibility by allowing web page creators to use their own set of mark-up tags. This feature can be used to make augmentations in the hypertext documents for the suitability of web crawling [14]. The crawlers designed in PARCAHYD project [13] and [16,17] aim to enhance the performance and quality issues of crawlers using the concept of augmented hypertext documents. For instance, to manage the volatile information, variable information of a document is marked through volatile tags [15], which in turn are extracted out from the document along with their associated volatile information. The tags and their contents are then stored in a file having same name as document but different extension (.TVI). The hypertext documents that support .TVI and other related augmentations [14-17] are known as Augmented hypertext documents.

1.2 Category Tree

A category tree [5] is used as a graphical user interface in the search engine. It is a pre-defined canonical topic taxonomy with example keywords. To run a specific instance, initial input has to be provided in two forms. The user has to select and/or refine specific topic nodes in the taxonomy, and may also need to provide additional example keywords. The user, then, selects the example keywords of his interest in corresponding topic or category node. Subsequently, these selections are submitted to the search engine.

3. THE PROPOSED DESIGN OF CONTEXT DRIVEN FOCUSED CRAWLER(CDFC)

```
head>
<title>Crawler Information</title>
<meta name = “context” content = “General
Information about Crawler”/>
<meta name = “keywords” content = “Crawler, Web
pages, Search Engine, Spiders, Wanderers, Worms”/>
</head>
<body>
A Crawler is a program that retrieves web pages commonly
for use by a search engine. It traverses the web by
downloading the documents and following links from page
to page. Web crawlers are also known as spiders
</Keyword> or wanderers, or worms etc.
</body>
```

Figure 1: Sample Augmented XML cod

For the proposed work, the context of the required information has been augmented to the hypertext document wherein the tag names called ‘*keyword*’ and ‘*context*’ are explicitly marked at the time of creation of a hypertext document by the author. As an example, consider the sample XML code shown in Fig. 1.

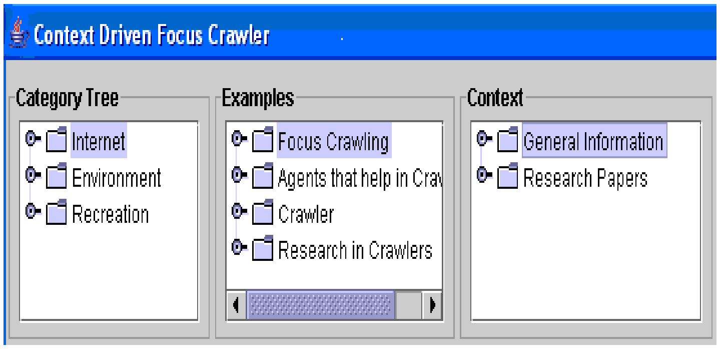


Figure 3: Modified Category tree for Search Engine

At the time of saving the document, all the keyword tags along with context and keyword tags are extracted out and stored separately in a file having same name but with different extension (say .TOC). The .TOC file extracted from the sample code of Fig. 1 is shown in Fig. 2. It may be noted that the TOC (Table of Contexts) is definitely much smaller in size as compared to the whole document.

Context	General Information about Crawler
Keywords	Crawler, Web pages, Search engine, Spiders, Wanderers, Worms

Figure 2: TOC file for sample code of Fig. 1

The category tree has also been suitably modified for the proposed design such that the context is also displayed with category examples. As shown in Fig. 3, the user selects a Category node (say Internet), and then its related examples are displayed. When the user selects an example (say Crawler), the two associated contexts are shown and finally, the user selects the context (say General information). In fact, the modified Category tree is a pre-specified collection of various categories in a graphical interface showing the various examples under these categories with their *contexts*. The user can choose any of the associated contexts by selecting *Category* → *Examples*→ *Contexts* in the order. Nevertheless, if needed, new examples can be inserted by the user, which may later on linked to the contexts by the crawler.

For the purpose of crawling the web, CDFC employs three agents namely User agent, Matcher agent and Dbase agent as listed in Table 1. A brief discussion on these agents and their related components is given below:

Agent	Responsibilities
User agent	Acts as interface between the user and the system. Accepts user selections of keywords and context and displays documents to the user
Matcher Agent	Matches the user keyword with the keywords in the database, retrieves their contexts & URLs and sends them to user agent
Dbase Agent	Acts as interface between database and the external world. Stores and updates the TOC files and documents downloaded by the crawler in the database

Table 1: Agents and their Responsibilities 1. User Agent:

The user agent is responsible for the following activities:

1. It accepts the user selection of category node and related keywords from the category tree. It sends this information to the matcher agent to retrieve the associated contexts and their links from the database.
2. On getting associated contexts and their links from matcher agent, the user agent displays the list of contexts to the user in category tree.
3. The user agent accepts the context selected by the user and displays its all corresponding links.
4. On selection of a link by the user, user agent sends this link to Retrieve_Doc_Process to retrieve the document from the database. If the document is not present in the database, it passes the link to crawler to download the document.

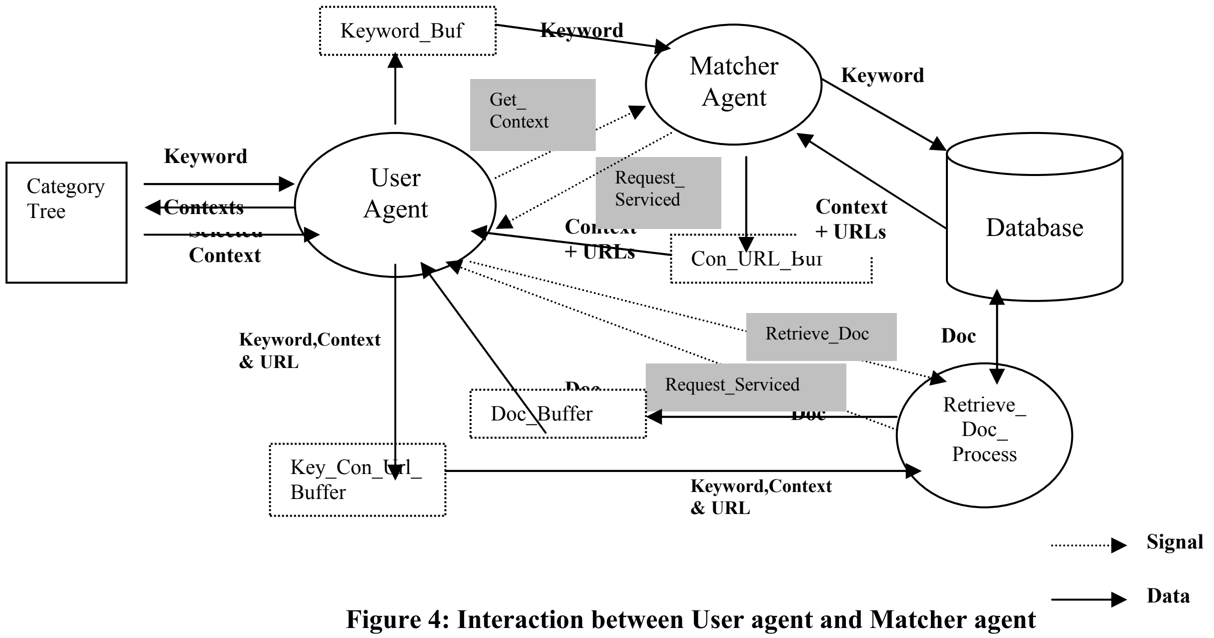


Figure 4: Interaction between User agent and Matcher agent

2. Matcher Agent: The matcher agent is responsible for the following activities:

1. It gets the keyword from the user agent and searches the keyword in the database to retrieve the corresponding contexts and their URLs.
2. The contexts and their associated URLs are sent to the user agent.

3. Retrieve_Doc_Process: It is responsible to search and retrieve the document in the database corresponding to a URL.

The interaction of user agent with matcher agent and Retrieve_Doc_Process (shown in Fig. 4) is described as follows:

1. It accepts the user selection of category node and related keywords from the category tree, stores the keyword in *Keyword_Buffer* and sends the message *Get_Context* to Matcher agent.
2. Matcher agent extracts the keyword from the Keyword_Buf and matches it with the keywords stored in the database. If the keyword is found, it retrieves its related contexts and URLs, stores them in *Con_URL_Buffer* and sends the signal *Request_Serviced* to the User agent.
3. User agent extracts the contexts and their URLs from the buffer and displays them to the user in category tree.
4. The user selects one of the contexts in the category tree. The user agent stores the keyword, its selected context and corresponding URL in *Key_Con_Buffer* and sends the message *Retrieve_Doc* to the Retrieve_Doc_Process.
5. Retrieve_Doc_Process extracts the keyword, context and URL from the buffer and searches the database for the document corresponding to URL. If the document is found, it stores the document in *Doc_Buffer* and sends the message *Request_Serviced* to the user agent.
6. The user agent extracts the document from the buffer and displays it to the user.

4. Dbase Agent: It is responsible for storing and updating the database whenever a new document or TOC is downloaded by the crawler.

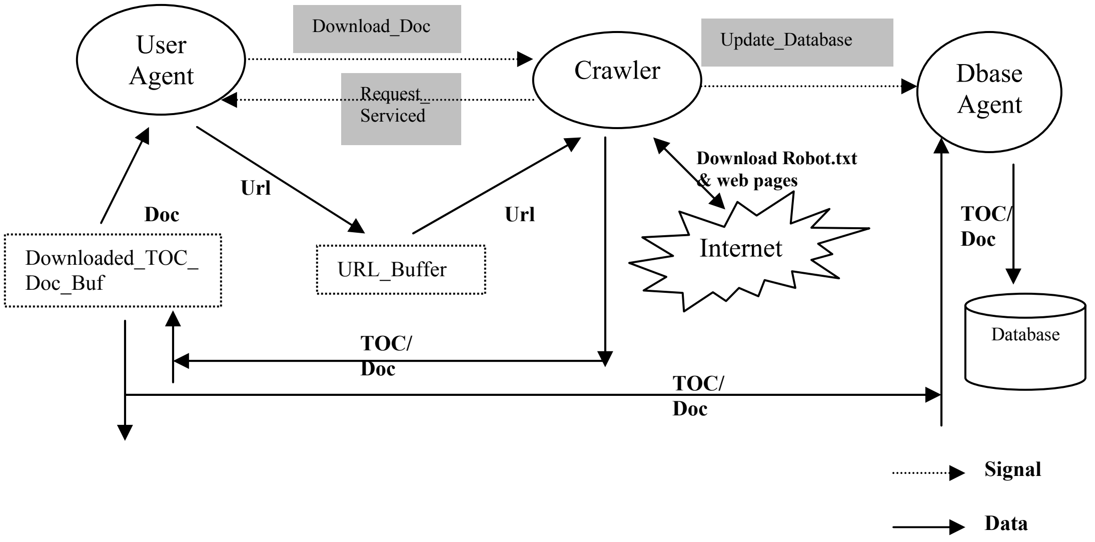


Figure 5: Interaction between User agent and Dbase agent & Crawler

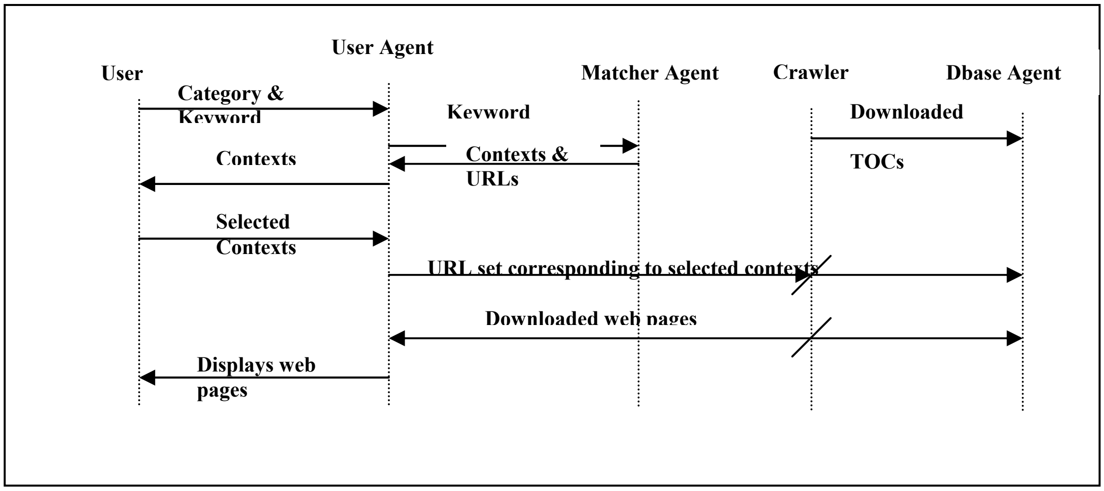


Figure 6: Sequence interaction diagram for various active components of CDFC

5. Crawler: The crawler continuously downloads the TOC files from the WWW in the background and stores them in the database. It also downloads the documents from the web on the request from the user agent and stores them in the database.

The interaction between the user agent, Dbase agent and crawler (shown in Fig. 5) is described as follows:

1. If the user agent needs a new document to be downloaded by the crawler, it stores the URL of that document in the *URL_Buffer* and sends the message *Download_Doc* to the crawler.
2. The crawler extracts the Url from the buffer and downloads the document from the web. Thereafter, it stores downloaded document in the *Downloaded_TOC_Doc_Buffer* and sends the message *Request_Serviced* to the user agent. Simultaneously it sends the message *Update_Database* to the Dbase agent to store the downloaded document in the database.
3. The user agent extracts the document from the buffer and displays it to the user.
4. Dbase agent also extracts the document from the buffer and stores it in the database.

The interactions between various active components of CDFC have been shown along the time line axis (see Fig. 6).

4. PERFORMANCE BENEFITS

The performance benefits of the proposed crawler are evaluated based on the following parameters:

1. **Harverst Ratio**: It is the rate at which relevant pages are acquired and how effectively irrelevant pages are filtered off. Since all web pages are retrieved according to the context selected by the user in CDFC, number of irrelevant pages is almost zero. Thus, the harvest ratio is high.
2. **Precision**: It is the ratio of number of relevant pages to the number of acquired pages. This is also high in CDFC as almost all pages are relevant to the user.
3. **Storage Requirements**: In CDFC, no document is downloaded if the user has not requested it. Therefore it does not index the documents which will never be used. Moreover the number of documents downloaded is very less in number as only related web pages are downloaded. Thus, storage requirement is very less as compared to other conventional crawlers.
4. **Search Time**: Since, the database size is very less in CDFC; it does not take much time to present the search results to the user.
5. **Network Traffic**: Since only highly related web pages are downloaded, which are very less in number and the size of TOC files being very less (5% of the original document), a significant amount of network traffic is reduced in CDFC.

Thus, the proposed crawler presents a flexible and interactive user interface in the form of category tree so that the user is guided in selecting the proper keywords along with their contexts for the web search. CDFC downloads only highly related documents, which are very less in number, thereby reducing the problem of information overkill faced by the user. Moreover, network traffic is reduced, as irrelevant web pages are not download

5. CONCLUSION

The proposed design of context driven focused crawler (CDFC) is based on the augmented hypertext document wherein the context of the keywords is stored in the form of TOC (Table of Contexts). The TOC coupled with a category tree provides context of the keywords. This design not only avoids the expensive complex computations for deriving the context of the user keywords but also reduces the network traffic significantly. Moreover, the quality of downloaded documents is in conformance with the topic and context of the user choice.

REFERENCES

[1] Raj Kamal -- *Internet and Web Technologies*; Tata McGraw Hill, 2003

[2] Junghoo Cho, Hector Garcia-Molina, " Parallel Crawlers", *Proceedings of the 11th International World Wide Web Conference*, Technical Report, UCLA Computer Science, 2002

[3] Mike Burner. , "Crawling towards eterneity: Building an archive of the worldwide web ", *Web Techniques Magazin*e, 2(5), May 1998.

[4] Martin Ester, Matthias Grob, Hans-Peter Kriegel, "Focused Web Crawling: A Generic Framework for specifying the user interest and for Adaptive crawling strategies*", Proc. of 27th International Conference on Very Large databases(VLDB '01*), 2001

[5] S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", *Proc. Of 8th International WWW conference*, Toronto, Canada, May,1999

[6] Diligenti M., Coetzee F.M., Lawrence S., Giles C.L., Gori M., "Focused Crawling using context graphs", *Proc. International Conference on Very Large Databases (VLDB '00*), 2000,pp. 527-534

[7] Yang Yongsheng, Wang Hui, "Implementation of Focused Crawler", *COMP630D Course Project Report*

[8] J.Kleinberg, "Authoritative Sources in a Hyperlinked Environment", *Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms*, San Francisco, California, USA, 1998.

[9] Junghoo Cho, Hector Garcia-Molina, L.Page, "Efficient crawling through URL ordering", *Proc. of 7th International WWW conference*, Brisbane, Australia, April, 1998

[10] Steve Lawrence, "Context in Web Search", *IEEE Data Engineering Bulletin*, Volume 23, Number 3, pp. 25-32, 2000

[11] S. Brin, L. Page, " The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Proc. of the 7th International World wide web Conference*, Brisbane, Australia, 1998.

[12] F. Crimmins, "Focused Crawling review", 2001

[13] A.K. Sharma, J.P. Gupta, D.P. Agarwal, "PARCAHYD: An architecture of Parallel Crawler based on augmented hypertext documents" *Communicated to IASTED Journal of Computers and Applications*, June 2005

[14] A.K. Sharma, J.P. Gupta, D.P. Agarwal, "Augmented Hypertext Documents suitable for parallel crawlers", *Proc. Of WITSA-2003, a National workshop on Information Technology Services and Applications*, Feb'2003, New Delhi

[15] A.K. Sharma, J.P. Gupta, D.P. Agarwal, "A Novel Approach towards Efficient management of Volatile Information", *Journal of Computer Society of India* (CSI), July-Sep. 2003

[16] A.K. Sharma, Naresh Chauhan, Amit Goel, "An Agent based Crawler for Management of Volatile Information on World Wide Web*", In Proceedings of National Conference on Communication & Computational techniques (NCCT '06)*, Dehradun, Feb. 2006

[17] A.K. Sharma, Naresh Chauhan, "Demand Crawling based Efficient Web Search for Mobile Clients", *In Proceedings of National Conference on Information & emerging Technologies*, Ropar, Punjab, Feb. 2006

[18] Naresh Chauhan, A.K. Sharma, "A Comparative Analysis of Focused Crawling Techniques", *Proceedings of National Conference on IT*, Panipat, March 2006

A Mobile Transaction System for Open Networks

R. B. Patel[1] and **Anu**[2]

Abstract - ***The evolution of mobile computing devices and wireless network has created a new mobile computing environment. Users equipped with portable devices can access, retrieve and process information while in mobility. Mobile devices like laptops; mobile phones have become more powerful data processing elements. Traditional transaction model has moved forwarding to mobile transaction system. Autonomous decentralized systems represent examples of environments for which the use of mobile codes is quite convenient. For example, designing highly scalable distributed systems in a massive, heterogeneous and multi organizational distributed environment seems to benefit much from mobile codes, given their ability to decentralize processing; to adapt to the autonomy of systems; to flexibly allow the management of installed code; and their support to the interaction with human users. This paper presents a hierarchical transaction model for the execution of distributed transactions with mobile code on open networks. The developed transaction model is built upon the concept for fault tolerance of mobile code based executions. The presented transaction model is an open nested transaction model. This model supports those parts of a distributed transaction which is executed asynchronously in relation to other parts of the same global transaction. Furthermore, the model is able to recover the execution of a transaction when a sub-transaction of this transaction becomes unavailable for a long period of time and the results of a comparison of developed model, with some existing ones, are also reported. We have also suggested and implemented an efficient naming and locating mechanism for tracing/finding the status of a transaction whenever fault(s) arises in the transaction processing system/network or processing of a sub-transaction is delayed.***

Index Terms - MH, Transaction, ACID, TPS.

1. INTRODUCTION

Technological advancements in networking and distributed processing are enabling the emergence of new types of distributed processing environments. Exemplified by electronic service markets or virtual enterprises, such environments are highly complexed distributed systems that support corporations needs for integrating systems and that allow new forms of automated cooperation. Many types of mobile computing devices such as laptops, personal digital assistants (PDA) are available. The capacities of these mobile devices have become more powerful. They have more processing speed and longer operating time. Mobile computing devices are becoming the major work processing equipments in daily activity. Combining with the expanding of the high-speed network like the Internet, mobile computing applications are growing rapidly. Some of the characteristics of such systems are: they are composed by a multitude of autonomous organizations cooperating or competing to achieve their own goals; massive geographical distribution; they encompass a huge diversity of types (qualities) of communication links; the execution of inter-organizational activities are typical in such environments; a multitude of services are offered to a multitude of clients of such services; different type of services exist and they may range from totally automated services to services executed by human beings, high dynamism with no global control, high heterogeneity and coexistence of different types of hosts such as laptops, personal computers, powerful workstations or mainframes). Environments with these properties will be called here open networks, i.e., Internet.

A transaction is a collection of operations on the physical and abstract Application State [1, 28, 29]. Transaction processing systems provides a means to record all states and effects of execution of program in computing resources. Transaction not only relates to operation on database system but also involves in many daily applications upon many computing resources like telephone call, email system, flight reservation. Mobile transaction is more complicated than conventional transaction in both of design and execution states. When a mobile host (MH) moves from one region to another, many computing activities like establish new communication channel, forward the state of transaction to new base host (BH) [15] are involving. The execution of mobile transactions is not only unpredictable in time but also depends on their location.

A computation processing is considered as a transaction or conventional transaction if it satisfies ACID [1] [2] (Atomicity, Consistency, Isolation, and Durability) properties.

1. Atomicity: an executable program, assumed that this program will finally terminate, has one initial state and one final state. If it appears to the outside world that this program is only at one of these two states then this program satisfy the atomicity property. If there are intermediate results or message needs to be displayed, then they are not displayed or they are displayed in final state of the program. If the program achieves its final state it is said to be committed, otherwise if it is at the initial state after some execution steps then it is aborted or rollback.
2. Consistency: if a program produces consistent result only then it satisfies the consistency property and it will be at the final state or committed. If the result is not consistent then a transaction program should be at the initial state, in other word the transaction is aborted.
3. Isolation: if a program is executing and if it is only process on the system then it satisfies the isolation property. If

[1,2]*Dept of Computer Engineering, M. M. Engineering College, Mullana-133203, Haryana, INDIA*
E-Mail: [1]*patel_r_b@indiatimes.com and* [2]*anurawal2k@yahoo.com*

there are several other processes on the system, then none of the intermediate state of this program is viewable until it reaches its final state.

4. Durability: if a program reaches to its final state and the result is made available to the outside world then this result is made permanent. Even a system failure cannot change this result. In other words, when a transaction commits its state is durable.

A mobile transaction is a set of relatively independent (component) transactions which can interleave in any way with other mobile transactions. A component transaction can be further decomposed into other component transactions, and thus mobile transactions can support an arbitrary level of nesting.

Let us assume that S is a two level mobile transaction that has N component transactions, $T_1, \ldots, T_N$. Some of the components are compensatable; each such T_J has a compensating transaction cmst_T_J that semantically undoes the effects of T_J but doesn't necessarily restore the database to the state that existed when T_J started executing. Component transactions can commit without waiting for any other component or *S* to commit, i.e., component transactions may decide to commit or abort unilaterally. However, if *S* aborts, a component transaction that has not yet committed is aborted.

Mobile transactions, components or otherwise, are distinguished into 4 types:

a) Atomic transaction: these are associated with the significant events {Begin, Commit, Abort} having the standard abort and commit properties. Compensatable and compensating transactions are atomic transactions with structure- induced inter-transaction dependencies. A compensatable component of S is a component of which can commit its operations even before S commits, but if S subsequently aborts, the compensating transaction cmst_T_J of the committed component T_J must commit. Compensating transactions need to observe a state consistent with the effects of their corresponding components and hence, compensating transaction must execute (and commit) in the reverse order of the commitment of their corresponding components.

b) Non- compensatable transactions: these are component transactions that are not associated with a compensating transaction. Non- compensatable transactions can commit at any time, but since they cannot be compensated, they are not allowed to commit their effects on objects when they commit. Non- compensatable transactions are structured as sub transactions (as in nested transaction) which at commit time delegate all the operations that they have invoked to *S*.

c) Reporting transactions: a reporting component T_J can share its partial results with *S*, i.e., a reporting component delegating some of its results at nay point during its execution. Whether or not a reporting component delegates all the operations not previously reported to *S* when it commits depends on whether or not it is associated with a compensating transaction.

d) Co-transactions: these components are reporting transactions that behave like co-routines in which control is passed from one transaction to another at the time of sharing of the partial results, i.e., co- transactions are suspended at the time of delegation and they resume execution where they were previously suspended. Thus, as opposed to non-compensatable transactions, co-transactions retain their state across executions; and as opposed to reporting transactions, co- transactions cannot execute concurrently.

Compensatable and non- compensatable components can be further considered as a vital transaction in that S is allowed to commit only if its vital components commit. If a vital transaction aborts, S will be aborted. Transaction *S* can commit even if one of its non- vital components aborts but S has to wait for the non- vital components to commit or abort.

The simplest form of transaction is flat transaction. A flat transaction can be considered as a sequential correctness computer program. Every execution step is after one another. Flat transaction has many disadvantages. For example, it cannot support long transaction efficiently. If failure happens during its execution then it has to rollback to its initial state and wastes all useful computation. Nested transaction model is a more flexible transaction model. This model is a tree of transactions. A big transaction is refined into many smaller (flat) transactions called sub-transactions. These sub-transactions can execute concurrently in different processes in different processing hosts. The ACID properties are more relaxed in this nested transaction model. Autonomous decentralized systems represent examples of environments for which the use of mobile codes [20] is quite convenient. For example, designing highly scalable distributed systems in a massive, heterogeneous and multi organizational distributed environment seems to benefit much from mobile codes, given their ability to decentralize processing; to adapt to the autonomy of systems; to flexibly allow the management of installed code and their support to the interaction with human users.

In this paper we present a model for the execution of distributed transactions with mobile code on open networks. The developed transaction model is built upon the concept for fault tolerance of mobile code based executions [20]. The presented transaction model is an open nested transaction model. The model supports those parts of a distributed transaction which are executed asynchronously in relation to other parts of the same global transaction. Furthermore, the model is able to recover the execution of a transaction when a sub-transaction of this transaction becomes unavailable for a long period of time. Open nested transaction model has been proposed for coping with long running activities and with the autonomy of systems in multi databases and thus take into consideration aspects of open networks. We have also suggested and implemented an efficient naming and locating mechanism for tracing/finding the status of a transaction whenever fault(s) arises in the transaction processing system/network or processing of a sub-transaction is delayed.

The rest of the paper is organized as follows. Section 2 gives overview of the currently agreed mobile computing environment. Section 3 discusses some of the limitations of exiting transaction models. Section 4 describes issues in Mobile Transaction Processing. Section 5 presents System Model. Section 6 gives model of the Transaction Processing System (TPS). Transaction model is presented in Section 7 and Section 8 gives implementation and performance study compared with the existing one. Related works is presented in Section 9 and conclusion of this work is given in Section 10.

2. MOBILE COMPUTING ENVIRONMENT

It is important to identify and define the mobile computing environment. Based on that defined mobile environment, requirements as well as characteristics will be identified. Mobile computing environment includes: a wired network with fixed work-stations or fixed hosts (FH), mobile hosts (MH) and base host (BH) [15] which is similar to mobile support stations (MSS) [3] [4] [8] [9] as shown in Figure 1.

Connection between MH and BH is wireless network; this network has characteristics low bandwidth, error-prone and frequently disconnection. These characteristics are discussed in detail in the next subsection 2.1. BH and FH communicate with each other via reliable high speed connection networks, which can be wired or wireless network within limited range, such as inside a building. The BH is motionless. MHs can include broad types of mobile devices, typically laptop computers with high-speed modems. Works can be shared between MH and FH. The role of BH is not only as processing element but also it is acting as an interface to help MH getting contact with relevant FH.

Each BH is being responsible for all the tasks which occur in a region. One MH can only connect to one BH at any given time but at overlap region during the handoff it connects to two BHs. A MH is moving from one region to another when computation task is in processing, and sometimes MH requests to connect to a database or computing resource resided from a FH on fixed network. This work will be done with the help of BH. The BH will receive requests from MH, forward the requests to the responsible FH and return the answer from the FH to the MH. When a MH is leaving a region controlled by a BH, this BH will perform a handoff operation to transmit or forward all information related to this MH to next BH. The next BH in new region will be ready to support the MH.

Databases and other computing resources are stored on the FH or wired network, this environment allow mobile environment inherits from the current existing distributed computing environment. Normally, power supply and storage device limits MH computing capacity.

However, with the current technology, the power of mobile computers can last for several hours and the storage devices can store a large amount of data [5]. Then MHs can become major hosts for data processing.

The main features of mobile computing environment are communication, mobility, portability [6] and heterogeneity [3]. There are many research issues that are arising from these features.

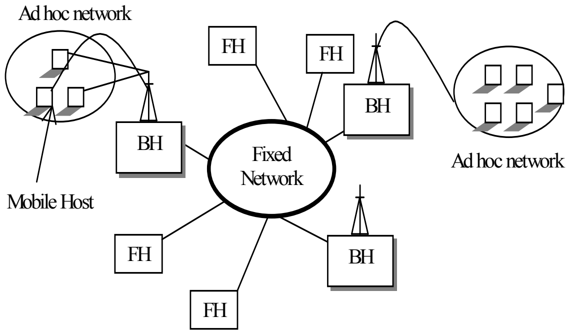


Figure 1: General Architecture of Mobile Computing Environment

2.1 Communication- MHs are connected with BH through wireless network. It is obvious that this wireless network does not have capacity as fixed wired network. First, the wireless bandwidth is very low, for example cellular network has bandwidth in the order of 10Kbps or wireless local area network has bandwidth of 10Mbps- 1000Mbps [5], Second, wireless network are having high error rate and frequent disconnection [3] [5], the same network data package may require retransmit many times. When MH is moving from one region to another, the current connection with BH will need to be changed to new connection. This process requires two steps: disconnecting from the current connection and establishing a new connection. The above disadvantages result in taking more time to transfer a same amount of data from the MH to FH and vice versa. Retransmit data causes unnecessary processing power, which is already very limited on the MH. The situation is more complicated if two MHs need to exchange data during cooperative task. Messages cannot be delivered directly between two MH but can be via one or more BH. Because of larger overhead in communication time, the longer time requires for MH to perform computation. Caching mechanism is currently the major method to ease the problem.

2.2 Mobility- Mobility is the most frequent activity of a MH. When MH is moving from one region to another in wireless network, the connection will need to be changed because one BH can only support MHs within its limited area. This cause frequently need of reconfiguration network topology and protocols. The more mobility causes the more time spends on reestablish communication between MH and BH. Because the activities of MH need support from its BH, therefore location management is another problem caused by the mobility of MH. MHs need to track BH in order to obtain data from the FH or other MH. In other hand, BH also needs to keep track on MH in order to transmit the result from the FH or to update the state of current MH profile. Mobility of MH raises the question on location dependent data [10]. The same query will have different results depending on the location of MH.

2.3 Portability- The availability of mobile devices depends on their power supply. A mobile phone can live up to five days but the laptop can only be for few hours. The more complicated application requires more processing power. Refining computation process into smaller partition (fined grain) or shifting heavy process from MH to FH for processing can save energy. Communication in MHs requires a lot of power. Compressing data or data distilling before transmission can reduce communication time. Caching also help MH tasks in disconnected period. Portability of MHs requires more sophisticated software applications. MH has smaller user interface like display screen, keyboard [6]. Many PDA support handwriting, therefore handwriting recognition software is required.

2.4 Heterogeneity- One BH needs to support broad types of mobile devices which operate in its region. Identifying what kind of hardware of the MH is important. Different MH requires different applications and data representations. When MH requests communication with other MH, the heterogeneous problem needs to be taken into account. How does BH solve this problem? A Composite Capabilities/Preference Profiles (CC/PP) can be used to provide a description of mobile device [14]. Different BH are in different heterogeneous network and these BH need to cooperate and communicate with each other for exchanging data. A standard interface is needed between BH. Java technologies or a middleware like CORBA [16] can be used to solve the heterogeneous problems.

3. LIMITATIONS OF EXISTING TRANSACTION MODELS

A computation that accesses shared data in a database is commonly structured as an automatic transaction in order to preserve data consistency in the presence of concurrency and failures. However, a mobile computation that accesses shared data cannot be structured using atomic transaction. This is because atomic transactions are assumed to execute in isolation that prevents them from splitting their computation and sharing their states and partial results. As mentioned above, practical considerations unique to mobile computing require computations on a MH to be supported by a MSS for both communication and computation purposes. This means that a mobile computation needs to be structured as a set of transactions some of which execute on the MSS.

In addition, Mobile computations are expected to be lengthy due first to the mobility of both data sources and data consumers, and second to their interactive nature, i.e., pause for input from the user. Thus, another requirement of mobile computations that atomic transactions cannot satisfy is the ability to handle partial failures and provide different recovery strategies, thus minimizing the effects of failures.

Nested transactions [21], where a (parent) transaction spawns (child) transactions, provide some more flexibility than atomic transactions in supporting both splitting of their computation and partial failures. However, nested transactions do not share their partial results while they execute. Nested transactions support procedure-call semantics and commit in a bottom-up manner through the root, i.e., when a child transaction commits, the objects modified by it are made accessible to its parent transaction while the effects on the objects are made permanent in a database only when the root transaction commits. This also means that the state of the mobile computations must be retained until the root transaction completes its execution. Consider the case in which the root executes on the MH whereas the child transactions execute on the MSS. If sub transactions do not retain their state after completing their execution, then the state of the whole computation needs to be maintained at all times on the MH in spite of its limited resources. On the other hand if sub transactions retain their state, the state of the computation is spread among MSS along the path of the MH making atomic commitment expensive.

Open- nested transactions such as Sagas [24], Split transactions[22] and Multi-transactions[23] relax some of the restrictions of nested transactions by supporting adaptive recovery, i.e. , allowing their partial results be visible outside a transaction. This is because, in open nested model, component transactions may decide to commit or abort unilaterally. It is interesting to note that most open- nested transaction models have been proposed in the context of multi database systems. A mobile database environment can be viewed as a special multi database system with special requirements. For example, the notion of local autonomy in mobile environments is manifested in the ability of the MHs to continue to operate in an independent fashion when they are disconnected.

Yet two specific requirements of transactions in mobile environment cannot be satisfied by current open transaction models. First, the ability of transactions to share their partial results with each other while in execution, and second to maintain part of the state of a mobile computation on a MSS in a way that minimizes the communication delays between a MH and MSS.

4. ISSUES IN MOBILE TRANSACTION PROCESSING

Mobile transactions are long-lived, bound to many different types of mobile devices, involved in heterogeneous database and network and execution time is varying. This section focuses research challenges in mobile transaction mainly on mobile database, service handoff and scheduling.

4.1 Mobile Database

Currently, the mobile transaction is developed on the top of currently existing database system. Most of mobile transaction models are based on the earlier discussed mobile environment. In this environment, the database resides, replicated and distributed on the fixed hosts in wired network. However, the capacity of mobile computing device is expanding and a MH can become a host for data processing or a place to store the native data. In this case, the physical location of database system is changing. Identify the location of the MHs which stores the required data is one of the major issues in mobile database [5]. To obtain optimization on query processing, databases are replicated or fragmented in MH. Because of the disconnection and mobility of MH, maintaining data consistency between MH is more complicated. Location dependent data also needs to be considered.

4.2 Service Handoff

When a MH moves into a new region, a new BH is assigned to this MH. Information about current transaction state is saved and transferred from old BH to next BH. This operation sometimes is unnecessary because not all the time MH requires assistant. Figure 2 illustrates the situation. MH *M* is moving from region A to region C through region B. However, in region C the MH does not need any assistant from BH in region B. The information about transaction state should directly forward to BH in region C. This information package also includes the hardware profile of MH, context of application and environment. If this information is stored at MH then the MH can become an active element, which can initiate a connection when needed. The question is how a MH finds out what BH it should connect to. Currently, when a MH wants to exchange information with another MH then it has to rely on the support from at least one BH. How can one MH directly obtain communication channel with other MH?

4.3 Scheduling

Execution time of mobile transaction is varying. Mobile transaction can easily miss its required deadline due to its mobility and portability. It is not applicable in mobile transaction if a missing deadline transaction is always aborted. Missing deadline causes inconsistency in global state of transaction and blocks other transaction's execution. Enforcing technique like earliest-deadline-first [3] can be applied. Mobile transaction requires flexible scheduling mechanism. Scheduling a transaction in a FH is different from MH. Schedule in mobile transaction should take into account the mobility of MH in both location and time. MH should be able to reschedule its execution plan according to its physical state (power, communication bandwidth).

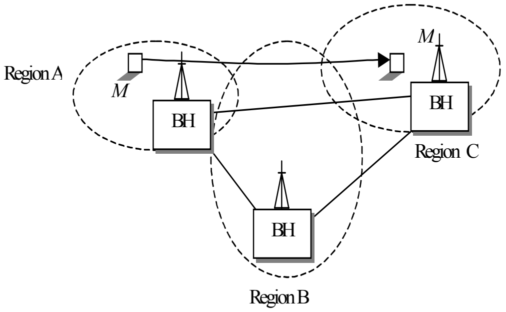


Figure 2. Service Handoff between BH

4.4 Caching

Caching of data at MUs can improve performance and facilitate disconnected operation. Much research has been performed in the area of MU caching [26]. Caching issues are complicated by the use of Location dependent data (LDD). Because of the fact that data which is cached can be viewed as a temporal replica of spatial data, as a MU moves into new data regions the cached data may become obsolete. This data is not stale because it is incorrect, but may not be desired because it is from a foreign region. Replacement policies need to be re-examined to include location information. For example, data from a foreign region should perhaps be replaced before data from the current home region even though the foreign data is more recently used. However, this is further complicated by the fact that ongoing or future queries could be bound to foreign regions. The MU mobility is such that the MU could very quickly move back into the home region for this data, making the re- placement policy also subject to movement of the MU. All of these issues are beyond the scope of this paper, but certainly need to be studied.

5. SYSTEM MODEL

A transaction submitted from a MH is called mobile transaction [3]. The MH, which issues transaction, and the MH, which received the result, can be different. For example, a user queries for a bus timetable from its laptop and requests the answer will send to mobile phone via SMS. A MH is a mobile computer which is capable of connecting to the fixed network via a wireless link. A FH is a computer in the fixed network which is not capable of connecting to a MH. A BH is capable of connecting with a MH and is equipped with a wireless interface. BHs, therefore, act as an interface between MH and FH. The wireless interface in the BHs typically uses wireless cellular networks because of the characteristics of mobile environment; mobile transaction has several additional requirements:

1. As MH has less processing capacity as FH, so mobile transaction should be able to split into a set of smaller transactions. These shorter sub-transactions can execute on FH or other MH. If possible, most of the computation on the MH should be shifted to FH for processing. When computing tasks are moving to FH, the FH have more computing power and shorter processing time. In addition, the computing resources are closer in FH. If the tasks require extra computing resources, wired network bandwidth is faster for resource allocating than wireless network. MH can save energy by disconnecting their connection while waiting for the results from the FH.
2. Mobile transaction has longer processing time or long-lived. Because of the communication overhead and frequent disconnection, the time required for exchanging needed data between MH and BH is longer. A part from this, MH has slower processing speed therefore a same transaction on MH will require longer time for completing than on the FH.
3. Mobile transaction should be executable when MH is in mobility and disconnected from the computing resources. It is not possible for MH staying connected all the time with the data resources. After the needed data has been caching into mobile storage device then MH can operate in autonomous mode. Data inconsistency in short time should be allowed. When the connection is established the new data item will be updated to the main database.
4. Mobile transactions require being able to operate in distributed heterogeneous environment. Different types of MH cooperate in mobile environment and different database systems are accessed during execution state of

mobile transaction. Mobile application should take into account the representation of data format in different system.

Mobile transaction is a collection of BH which traps a transaction in a region. For this purpose it runs a region manager (a transaction processing system in a region is called region manger), other nodes in this region may be mobile or fixed, i.e., network may be ad-hoc/ fixed. The BH works like a manager when a transaction leaves from a network domain, it may be managed by domain manager server (DMS). A DMS is basically a transaction processing system.

Some of the techniques developed in conventional transaction such as two phases commit (2PC) protocol, caching mechanism is needed to be extended or modified to be able to apply in mobile transaction. Another issue is to make the intermediate states of mobile transactions available to others. This will release locks on data item earlier and avoid blocking other transactions. DMS looks apart BH and BH looks apart a set of nodes lying in an inter-network, the system is a hierarchical.

6. TRANSACTION PROCESSING SYSTEM FOR DISTRIBUTED DATABASES

A *transaction processing system* (TPS) uses the transaction as a basic unit of task. It typically consists of a transaction manager, resource managers, and clients as shown in Figure 3. Client applications start particular transactions, within whose scope they forward data requests to registered resource managers, and commit or abort the transactions. *Resource managers* are entities, which store and manage data objects manipulated by transactions. They ensure durability of transactions. A database system is an example of a resource manager. The *transaction manager* enables clients to create, start, and finish transactions, monitors the lifecycle and distribution of executed transactions, and is responsible for ensuring the ACID properties of executed transactions.

For ensuring transaction's ACID properties the transaction manager employs two other entities that are usually not visible from clients – the lock manager and log manager. A *lock manager* is responsible for transaction isolation and achieves it by locking. The lock manager locks data objects if they are manipulated by a transaction. To ensure transaction isolation, all locks are held until the transaction is committed. Each resource manager usually has its private lock manager that manages transaction-aware locking, i.e., locks are associated with transactions (e.g., this is the case with traditional relational databases). To achieve atomicity and consistency, the transaction manager orchestrates *recovery* in case of a TPS *failur*e. A failure could be a crash of one of the participating hosts, a hard disk failure, a network disconnection, a power failure, or a software fault. Recovery is based on ensuring durability of the committed transactions' effects and discarding the effects of transactions that were being executed at the time of the failure and thus will be aborted.

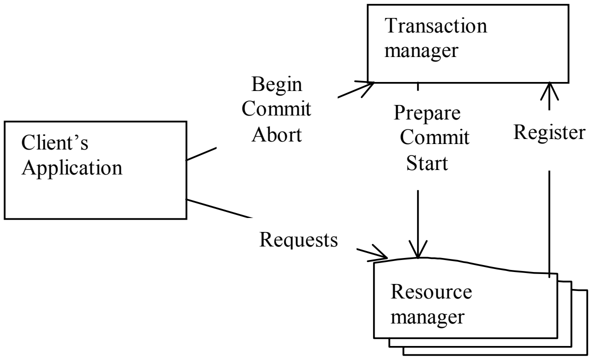


Figure 3: Basic Model of Transaction Processing System

Logging is the principal service that is used to support recovery. The *log manager* keeps track of every operation executed on behalf of a transaction. It writes the information needed for data recovery in case of transaction abort or TPS restart to a file called *log file* or simply *lo*g. It uses the following techniques for ensuring that the log is always in a consistent state: two copies of the log file are kept in persistent storage, the log is always written before data in a persistent storage is modified, all writes to the log are covered by appropriate locks and as part of the commit, the log is always written to the persistent storage (the *force-log-at-commit* rule). These techniques are usually combined with *checkpointin*g, which periodically writes the TPS state to a persistent storage to speed up the potential restart.

Since its parts are distributed over different nodes of the network, the TPS provides transaction-aware communication where connecting, authorization, and delivery of data requests take place. Every data request is executed on behalf of a particular transaction and is associated with the transaction identifier. We say that data requests are executed *in the context of a transaction* or simply in a *transaction contex*t. Several resource managers can be involved in a particular transaction. The transaction manager allows registering of particular resource managers with a transaction and manages transaction commit. Each resource manager executes its *local* transactions; each of them does not cross the resource manager's boundary. Since the resource manager is responsible for ensuring the ACID properties of its local transactions, it usually has its own private transaction manager, lock manager, and log manager. All participating transaction managers form a hierarchy, where every local transaction managed by the participating resource manager is associated with a *global* transaction managed by the topmost transaction manager, which is called the *root transaction manager* or *commit coordinato*r. A non-root transaction manager controls the local transactions executed on the corresponding resource manager, or it is responsible for coordinating transactions on the corresponding subtree of participating transaction managers is shown as shown in Figure 4.

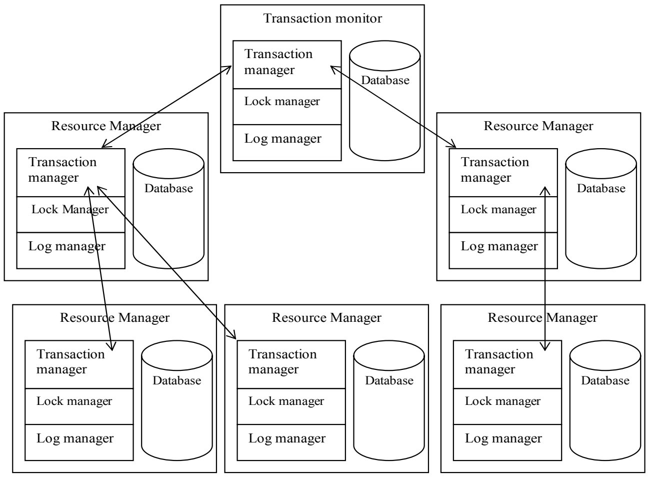


Figure 4: Architecture of Distributed TPS

In large distributed TPSs, a *transaction monitor* often plays the role of a central entity that controls global transactions. The transaction monitor (*TM*) is an application capable of integrating different heterogeneous TPSs and databases, and controlling several resources and terminals. The TM allows clients to initiate new transactions and to distribute transactions to several TPS on the network. They are also able to manage the ACID properties of executed transactions and, in particular, to perform recovery procedures. The TMs are designed to provide high reliability and performance. To achieve high throughput, they provide automatic load balancing [17], data request queuing, and other advanced features. If a TM is involved, clients never send their requests directly to the participating resource managers; instead, the transaction monitor mediates all the client's requests. The transaction manager of the TM often acts as the TPS commit coordinator.

The TPS is responsible for a negotiation protocol which guarantees that all effects of data requests executed on behalf of a transaction on registered resource managers are committed or aborted. In other words, all local transactions associated with a single global transaction are either committed or aborted. Usually, transaction managers support the *two-phase commit protocol.* In the first phase, the commit coordinator sends the *PREPARE* message to subordinate transaction managers. This is done recursively so that every transaction manager receives the *PREPARE* message. Each transaction manager votes *yes* or *no* indicating whether it is about to commit or abort. This is again provided in a hierarchical manner: a transaction manager coordinating its subtree's commit sends its vote message (vote for short) after it receives votes from all of its subordinate transaction managers. If all the transaction managers in the subtree are about to commit, then they vote *yes* and the subtree coordinator sends the *yes* vote to its parent commit coordinator. If any transaction manager in the subtree is about to abort, it sends *no* to the subtree coordinator, which then sends *no* to its parent coordinator. At the top level, if the commit coordinator receives *yes* from all of its subordinate transaction managers, it starts the second phase of the commit protocol by sending the *COMMIT* message to them. The message is then recursively sent to all the transaction managers and the transaction is committed. If the commit coordinator receives *no* from at least one of its subordinate transaction managers during the first phase of the commit protocol, it starts the second phase by sending the *ABORT* message to all of its subordinate transaction managers. The message is then propagated to all transaction managers in the hierarchy and the global transaction is aborted. The top-level transaction coordinator waits for messages acknowledging that all of the transaction managers have finished the second phase of the commit protocol.

Several optimizations of the two-phase commit protocol have been proposed in literature. For example, if a transaction is *read-only* (i.e., it does not provide any modifications of data objects), it can be committed in one phase. Advanced resource managers provide *heuristic decisions* on committing: a particular resource manager is able to heuristically commit or abort before the two-phase commit negotiation is completed. This can be efficient in situations when transaction managers have some advanced knowledge about the probability of commit or abort. If a particular transaction manager heuristically finishes a local transaction, and if his heuristic decision is wrong and does not correspond to the final vote of the global transaction, then the transaction manager has to provide an extra policy, which usually results in a human intervention. Several variants of the two-phase commit protocol is our next goal of this research that will support different communication topologies or to increase reliability in case of a commit coordinator or a participating transaction manager failure.

7. TRANSACTION MODEL

We have assumed that the open network environment is divided into network domains, regions (sub networks) and local sites of the clients as shown in Figure 5. The TPS are geographically distributed at different network domain, region and sites. There is a domain management server (DMS) in each network domain, which has information about all other DMS in the open network. One TPS running in a network domain considered as DMS. A transaction that is submitted to be performed over the open networks is called a global transaction. A global transaction is composed of a set of sub transactions. Each sub transaction may by its turn also contain sub transactions. The global transaction, therefore, has the form of a tree, called the transaction tree. The DMS of this tree is called the root transaction. The term transaction will be used hereafter to denote both the root transaction and sub transactions. Other common terms for hierarchical structures will also be used hereafter, such as client (leaf) transaction, parent transaction, etc.

The transaction running on the DMS is an open nested transaction. Each of the sub transactions of it can be either a flat ACID transaction or an open transaction. Open sub transactions of the DMS transaction have the same structure as the DMS transaction, thus applying the transaction structure recursively. Each of the flat transactions represents a client of the transaction tree.

Transaction running on the gateway corresponds to a combination of its sub transactions, forming a potentially complex control flow. The control flow of a transaction running on the gateway may include, for example, the specification of parallel and sequential execution of sub transactions, dynamic creation of sub transactions (instances) during the execution of a transaction and the definition of sets of alternative sub transactions (i.e., transactions that are equivalent, according to application semantics).

Each transaction has associated with it a set of input and output parameters, allowing a definition of data flow between transactions. Additionally, each transaction has a set of internal data which represents its private variables (its private state space).

The control flow of a transaction may be determined with the use of values of internal data and output parameters or outcome of previously executed transactions. The control flow is however, restricted in the basic transaction model so that for each transaction: (1) Open sub transactions can execute in parallel (2) The execution of flat sub transactions must be a sequence (3) No flat and open sub transactions can execute in parallel.

All transactions in this model are compensatable. Each transaction, with exception of the DMS transaction, has a corresponding compensating transaction. In case the effects of compensatable transaction must be cancelled after its commitment, its compensating transaction is executed. A compensating transaction cancels the effects of the compensated- for transaction according to application semantics. Compensation is performed in the reverse order of execution of the compensated - for transactions.

The compensating transaction of a local transaction of a client transaction is another flat transaction defined by the transaction specifier. The compensating transaction of an intermediary (i.e., transactions on the gateway) transaction corresponds to another open transaction that compensates the committed sub transactions of the compensated- for transaction. The compensating transaction for an intermediary transaction is defined automatically at runtime, depending on the sub transactions that have committed. Values for parameters of compensating transactions are defined by the application when the compensated- for transaction is committed or can be determined at the moment the compensating transaction executes.

DMS (root) also has information about all the regions in the network domain. DMS is responsible for maintaining uniqueness of names of regions, which are part of that network and helps to identify the region in which a transaction is present. Each DMS maintains a Domain Transaction Database (DTD), for information about the current location of all the transactions which were created in that domain or transited through it. Mobile transactions might have to split their computations into sets of operations, some of which operate on a MH while others on a FH. Frequent disconnection and mobility results in mobile transactions sharing their states and partial results violating the principle of atomicity and isolation which is traditional problem in existing transaction models. Mobile transactions require computations and communications to be supported by FH. Transaction execution may have to be migrated to a FH if disconnection is predicted in order to prevent the transaction from being aborted. The DMS behaves like a proxy and executes the transaction on behalf of the disconnected MH. The MH may either fully delegate authority to the DMS to commit or abort the transaction as it sees fit or may partially delegate authority, in which case the final decision to commit or abort the transaction would be made by the MH upon reconnection.

Each entry of DTD of the form $\left(T_x, FD, r\right)$ represents that transaction T_x can be found in region r of the foreign network domain FD (foreign Network domain), or it has transited from that network domain or region r. For DTD and RTD (Region Transaction Database), the primary key is the transaction name T_x. With the help of these naming schemes we check the fault tolerance by maintaining the status report of mobile transaction which keeps the updated information of all the transactions. Transaction is migrated from one network domain to another through the DMS. During inter domain migration the transaction has to update location information in the DTD of the present domain and register in the DTD of the target network domain. Every region maintains information about all TPS that are part of that region. A TPS can be a member of an existing region or can start in a new region. In each region, a RTD is present at a TPS which runs at the gateway of a sub network. It contains location information about each mobile transaction that was created in that region or transited through it. This host acts, as the Transaction Name Server (TNS) [25], which manages the RTD. TPS is responsible for maintaining uniqueness of names of all transactions, created in that region. Generally a transaction name comprises of User Assigned Name, Birth Host, Region, and Network Domain. When a new transaction is created, the user assigns a name to it by registering in the RTD of its birth region. Each entry of RTD of the form $\left(T_x, r, Nil\right)$ represents the region r where transaction T_x was found or transited through it. Similarly $\left(T_x, NIL, TPS\right)$ represents transaction T_x, which exists in that region at TPS. For intra region migration, it has to update its location information in the RTD of that region. This is an Intra Region Location Update. During inter region migration, the transaction has to update the location information in the RTD of present region and register in the RTD of the target region, specifying the host in that region to which it is migrating.

DAD Tuple	**Meaning**	**RAD Tuple**	**Meaning**
(T_x, FD, r)	Transaction T_x is in region r of	(T_x,r, NIL)	Transaction T_x is in present network

DAD Tuple	Meaning	RAD Tuple	Meaning
	the foreign network domain (FD)		domain and in region r
		(T_x NIL,TPS)	Transaction is in present network domain and region at site TPS

Table 1. DTD and RTD Tuples

Any location protocol for mobile transactions deals with three aspects: **name binding**, **migration** and **location,** each related to a particular phase in the transaction's lifetime. We have defined four atomic operations which are incorporated on **DTD** and **RTD**

1. **bind** operation is performed when a name is assigned to a newly created transaction, whose birth location is also stored. This operation causes the insertion of a new tuple in the database. As the transaction name has to be unique, this operation fails if a tuple with the same name already exists in the database.
2. **newloc** operation is performed when the transaction changes its location, by migrating to a new one. This operation updates the tuple already present in the database.
3. **find** operation is performed when a transaction has to be located in order to interact with it. For a given transaction name, this operation returns the current location of the transaction.
4. **unbind** operation is performed when a transaction name is no longer used (i.e., the transaction has been disposed off). This operation causes the deletion of the relative tuple from the database.

Since locating the transaction requires following a long path before reaching it. It follows a part of the link the transaction has left on the registers of the visited region and the network and parent domains. The updating operations performed during the *migration* phase are designed in order to shorten this path, thus increasing interaction efficiency and reducing the overhead. The steps for locating a target transaction T_i are as follows:

1. Extract birth network domain and birth region name from T_i .

   Domain_name ← Birth_Domain_Name;
   Region_Name ← Birth_Region_Name;
2. Contact relative DMS.
3. If query to DMS results in a tuple $\left(T_i, FD_i, R_i\right)$ {target transaction is not in that domain}

   Domain_Name ← FD_i;

   Region_Name ← R_i ;

   Get the domain name from the tuple and go to Step 2.
4. Else contact relative RTD //Transaction exists in the given Region_Name
5. Get the query result tuple $\left(T_i, R_i, TPS_i\right)$

   Region_Name ← R_i ;

   TPS_Name ← TPS_i ;
6. If R_i is *Nil* target transaction is located at TPS_i , else go to Step 4.

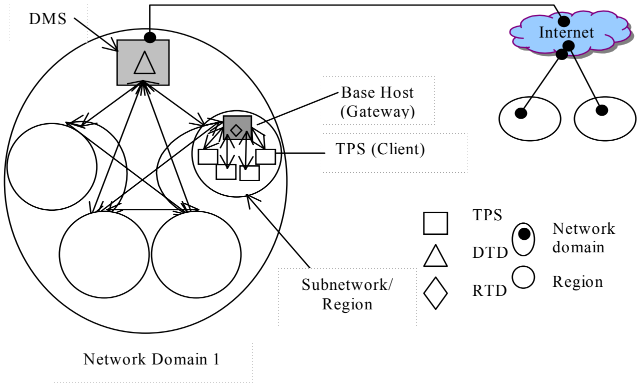


Figure 5: A Hierarchical Mobile Transaction Model for Open Network.

It is up to the binding and migration phases to maintain consistency of location information in the databases in such a way as to always allow transaction finding (unless a system or network crash occurs). When the network domain, in which the transaction is present, is found, the DTD is locked. Similarly the RTD is locked when the region is traced. The lock is reset only if the transaction does not reside in that region. It should be noted that keeping the RTD locked prevents the transaction from further migration so that communication with it is possible. This is required when direct or synchronous transaction - transaction communication is needed. For asynchronous transaction - transaction communication a message is dropped in the mailbox at the gateway/DMS and the transaction receives this message when it wants. Other possible case is drop and delayed transaction- transaction direct communication. In this technique transaction multicast message to all the gateways of a network domain and when it finds acknowledgement that transaction is found in particular region. The transaction waits for the message-receiving transaction to contact this transaction for making the dialogue.

Each transaction is either vital or non-vital. A vital transaction is a transaction the failure of which determines immediately the failure of its parent transaction. A failure of a non-vital transaction does not have direct effects on the outcome of its parent transaction.

Each client transaction is restricted to be executed entirely at the same TPS, i.e., only service components at the same TPS are accessed as part of a client transaction. The control flow of a client transaction represents a combination of accesses to services at that TPS. A compensating transaction for a client transaction is considered to be executed at the same TPS where the compensated- for transaction executed.

The general recovery semantics of the basic transaction model is as follows. In the occurrences of failures the recovery process of a transaction tries to perform forward recovery. A recovery process is performed which resets the execution to a consistent state and the transaction continues to be executed from that state on, trying to achieve a successful termination state. Backward recovery i.e., the cancellation of the effects of a transaction, however, may also occur. Backward recovery is performed when a vital transaction aborts. In this case the parent transaction of the vital transaction will be backward recovery.

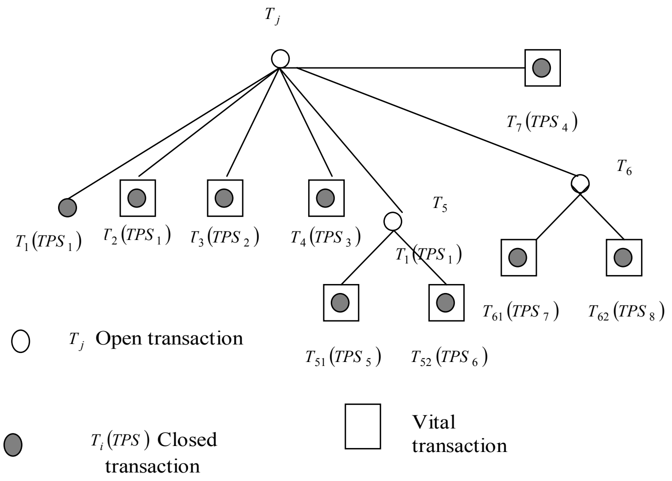


Figure 6: Flow of Transaction based on the Transaction model shown in Figure 5.

Due to the behavior of the fault tolerance mechanism, upon which this transaction model is based, partial backward recovery may also occur. In this case some of the already committed sub transactions of an open transaction are compensated as a form of back tracking the execution to a previous consistent state. Forward execution of the transaction is then performed from that state on. When the copy of the transaction at a TPS cancels the effects it produced after having stored the checkpoint. The basic transaction model enforces semantic atomicity.

Figure 6 shows an example basic transaction. In this transaction, the DMS transaction (T_j) has 7-sub transactions, denoted T_1 to T_7. Transactions T_1, T_2, T_3, T_4, and T_7 are closed. Transactions T_5 and T_6 are open. Transactions T_1 and T_2 should be executed, respectively, at TPS_1. Transaction T_3, T_4, and T_7 should be executed, respectively, at TPSs TPS_2, TPS_3 and TPS_4. The open transaction T_5 has two closed sub transactions, T_{51} and T_{52}, to be executed, respectively, at TPS_5 and TPS_6. Similarly open transaction T_6 has two closed sub transactions T_{61} and T_{62}, to be executed, respectively, at TPS_7 and TPS_8. Transactions T_2, T_3, T_4, T_7 and all the sub transactions of T_5 and T_6 are vital. Each transaction is either vital or non-vital. A vital transaction is a transaction the failure of which determines immediately the failure of its parent transaction. If any of them fails, its parent transaction must be backward recovered. A failure of a non-vital transaction does not have direct effects on the outcome of its parent transaction.

Figure 7 shows the control flow defined for the open transaction T_j. As shown in Figure 7, T_2 will be executed if T_1 succeeds. Transaction T_3 is executed if T_2 succeeds. Transaction T_4 is executed if T_1 fails. Transaction T_5 and T_6 are executed in parallel, after either T_3 or T_4 succeeds. Transaction T_7 will execute after T_5 and T_6 terminate. Similar definitions of control flow are supposed to exist for open transactions T_5 and T_6.

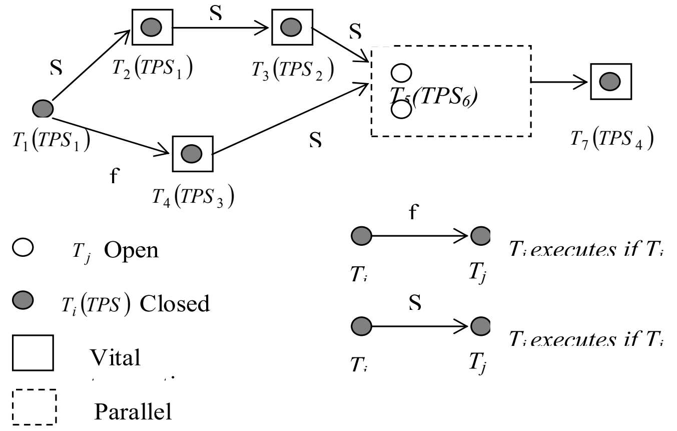


Figure 7: A Representation of control flow for root transaction of Figure 6

8. IMPLEMENTATION

To study the performance of the model suggested in section 6 we have implemented it on 10/100/1000 Mbps switched LAN that connects 850 workstations and personal computers, and is used by about 500 researchers and students. Machines are grouped into eight different networks with their own servers and servers of each network are connected to the main server of the institute. For each network there are 100 nodes which are running TPS, three mobile stations running TPS (DMS). These DMS are running mobile codes for finding the status of the different sub-transactions in different networks whenever a failure is arise. Mobile codes are implemented on PMADE [20]. We have implemented the transaction for computing the prime numbers (between 1 and 9999999) on a cluster of PCs (P-4, 3 GHz machines) using PMADE and j2sdk1.5.1

8.1 Performance Study

Figure 8 shows the system throughput of two approaches (developed and Kangaroo Model [9]). The throughput of the developed scheme is close to the Kangaroo Model in all case. As the developed model is implemented in Java, the high execution overhead of Java program results in the lower throughput when number of the TPSs very high. The real overhead generated due to DMS (root) controller of the sub-transactions which monitoring the status of them. The DMS launches a mobile code in case of a failure arise on any TPS for recovering the failed sub-transaction.

Figure 9 compares the system throughputs of the developed system with kangaroo model when temporary faults arising at different servers randomly. The result shows that the developed scheme can obviously improve the system throughput when increasing the number of TPS (servers). In the latter case, the processing capacities of the TPSs are wasted and no improvement.

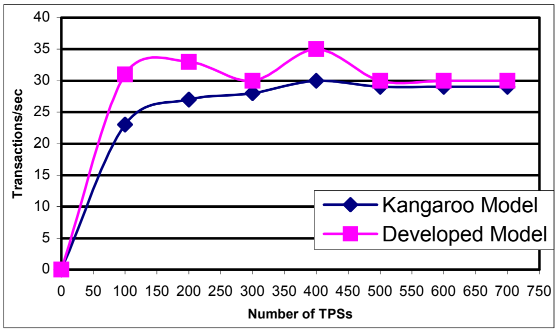


Figure 8: System throughput

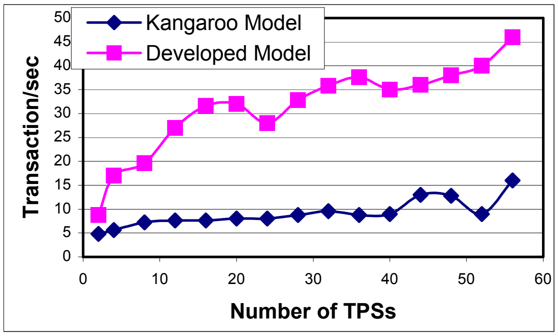


Figure 9: System throughputs of the developed scheme and the Kangaroo model when temporary faults are arising on different TPSs

8.2 Comparison With Existing System

Because in normal kangaroo transaction model (KT), three potential problems arise:

1. Resource blocking for other relatively smaller transactions initiated by the same user while main long lived transaction (LLT) is waiting for inputs. This is because most of the available commercial DBMS packages use conventional two-phase locking protocol [27].
2. Even if resource blocking doesn't occur due to usage of independent database resources by transactions, separate kangaroo transactions has to be initiated in each case the user initiates even a small transaction while the LLT is running. This leads to wastage of BS server resources.
3. Failure of a global transaction in a Joey in compensating mode results in abortion of the entire KT. Thus even if some transactions are there which are short and not involved in the failure, they will get aborted unnecessarily.

Transaction Model	**Atomicity**	**Consistency**	**Isolation**	**Durability**	***Execute In***
Kangaroo	May be	No	No	No	*Fixed Network*
Developed	*Yes*	*Yes*	*Yes*	*Yes*	*Both Fixed and Mobile Ad hoc Networks*

Table 2: Comparison between Developed Model and Kangaroo Transaction Model

9. RELATED WORKS

This section will review the transaction-processing concept and discuss on mobile transaction. Location and time of MH are the two major factors that effect on mobile transaction processing. This section outlines three mobile transaction models, which focus on mobility of MH.

Moflex transaction model [8] allows to model mobile transaction with extra information such as location, time and the precondition of mobile transactions. The sub-transaction *T* of mobile transaction *M* can be executable only when its external precondition predicated is satisfied. Moflex takes into account which sub-transactions are location-dependent.

Pre-write transaction model [4] allows a transaction on a mobile host to submit a pre-commit state and the rest of the transaction can be carried out at the fixed or other mobile hosts at later time. The main point is making all the updated data items visible to other transactions. This model can be use to support mobile hosts which have little power for processing data. Pre-commit transaction model eases the locking on data record and avoid longer time blocking other transactions. However it is not carefully taking into account the risk of frequent disconnecting and higher error rates of wireless data transmission.

Kangaroo transaction model [9] is developed beside the existing multi-database environment. Kangaroo mobile transaction does not start and end at the same host. In this model, mobile transaction hops through stationary hosts in wired network. The whole transaction and related information are pushing forward to the final committed host. Kangaroo model is supported by the autonomy of local DBMS. Kangaroo is one model that captures the movement nature of mobile unit.

Recovery from long term failures of the nodes from where a transaction is being controlled and mobility of the control flow of a transaction execution were also considered in the development of two transaction models, respectively, in the transaction model of ConTracts[11] and in migrating transactions[12]. In the ConTracts, if the node from where a ConTract is being executed fails, it can be re-instantiated at another node. A ConTract, however, does not move during its execution. In migrating transactions, the flow of control of a transaction migrates in a distributed environment. Executing transactions with mobile codes extends the notion, providing more flexibility for the distribution of code and for the movement of the transaction control flow in the environment.

In [13] the fault tolerance protocol and the transaction model presented here are described in details. In this model aspects of the presented approach are further discussed, such as: extensions to the basic transaction model; replication policies considering the availability properties of agencies; how autonomy of system is supported by the model; among others. In [18] a concept is presented for executing open and closed nested transactions with multiple mobile agents. The paper, however does not consider long-term failures. In [19] a model for executing transactions with a single mobile agent is presented. The transaction model presented supports compenstable and non- compenstable transactions and the specification of so-called ACID groups. An ACID group is a combination of sub transactions that is executed isolated from other parts of the same transaction and from other parts of the same transaction and from other agent-based transactions. The model supports that ACID groups or the set of non-compenstable transactions span more than a single agency. In this paper the execution of distributed transactions can be based on more than a single mobile agent. Additionally, it is not allowed here that isolated parts of an agent-based transaction span more than one agency, in order to facilitate recovery from long term failures.

The developed hierarchical mobile transaction model is fault tolerance in case of temporary failures arise on the transaction execution servers and gives better performance than Kangaroo model.

10. CONCLUSION AND FUTURE WORKS

In this paper we have presented a hierarchical mobile transaction system. This transaction model is based on mobile codes that take into considerations properties and requirements of open networks and their applications. The model represents a concept that integrates the mobility of the codes with the execution of control flows with transaction semantics. This transaction can be used as an approach for providing reliability and correctness of distributed activity in the open networks that provides the benefits of mobile codes. The resulting concept exhibits important features that should be supported by an underlying infrastructure to fulfill requirements of applications running in open networks.

The effectiveness of the applicability of mobile codes to open environments is, however, subjected to or influenced by the development of appropriate solutions to a set of issues. The main set of such issues realties to what can be called controllability of mobile code based activities. Other aspects are security, accounting and testing. The scope of applicability of mobile codes will be dependent on the achievements reached to these issues. The described model represents a step towards developing controllable mobile code based activities. This model is currently being extended to incorporate more functionality and to decrease some of the implied costs.

The mobile computing environment can support MHs to perform mobile transaction. Users can easily manipulate information despite of their location and what mobile devices they have. However, the disadvantage of this environment is that it cannot provide flexible way to exchange data between MH. One BH responds for supporting all MH in its region, this can cause a bottleneck when there are many MH in the same region and single failure mode if this BH fails. Current mobile transaction models are based on existing database systems. The models along with the characteristics of mobile environment help to analyze the requirement of mobile applications. The challenge is that when every host in mobile environment is MH. Different variants of the two-phase commit protocol is our next goal of this research that will support different communication topologies or to increase reliability in case of a commit coordinator or a participating transaction manager failure

REFERENCES

[1] Jim Gray, Andreas Reuter (1993), Transaction Processing: Concepts and Techniques, Morgan Kaufmann Publishers, Inc.

[2] George Coulouris, Jean Dollimore, Tim Kindberg (2001), Distributed Systems: Concepts and Design. Addison-Wesley.

[3] V. K. Murthy (2001), Seamless Mobile Transaction Processing: Models, Protocols and Software Tools, in Proceedings of the Eight International Conference on Parallel and Distributed Systems (ICPADS 2001), 26-29 June 2001, KyongJu City, Korea, pp. 147-156.

[4] Madria, S.K., Bhargava, B (1998), A Transaction Model for Mobile Computing, in Proceedings of the International Conference on Database Engineering and Applications Symposium (IDEAS'98), Cardiff, Wales, U.K., July 08 - 10, 1998, pp. 92-102.

[5] M. Tamer Ozsu, Patrick Valduriez (1999), Principles of Distributed Database Systems Prentice Hall.

[6] George H. Forman, John Zahorjan (1994), The Challenges of Mobile Computing, 27(4): 38 – 47, April.

[7] Karen Furst, William W. Lang and Daniel E. Nolle, See Furst, Karen, William W. Lang, and Daniel E. Nolle (1998), Technological Innovation in Banking and Payments: Industry Trends and Implications for Banks, Quarterly Journal, Office of the Comptroller of the Currency, Vol. 17, No. 3, pp. 28, Sept.

[8] Kyong-I Ku; Yoo-Sung Kim (2000), Moflex transaction model for mobile heterogeneous multidatabase systems, in Proceedings of the 10th International Workshop on Research Issues in Data Engineering (RIDE 2000), San Diego, CA, USA. Feb.28-29, 2000, pp. 39 –45.

[9] Margaret H. Dunham, Abdelsalam Helal, Santosh Balakrishnan (1997), A Mobile Transaction Model that Captures Both the Data and Movement behavior, Mobile Networks and Applications, 2, pp. 149–162.

Continued on page no. 84

Computer and Internet use among Families: A Case of Botswana

Dr. Rama Srivastava[1] and **Ishaan Srivastava**[2]

Abstract - ***During the past 20 years, novel communication technology devices have become familiar in African homes; among them are Personal Computers and the Internet. Social reviewers and other polemicists have debated whether these devices influence the lives of families in a productive or a destructive way. The authors examined the literature about family use of Computers and the Internet. Though home Internet access in Botswana is constantly increasing, there is diminutive information available about actual usage patterns in homes. The present study was carried out across Botswana on 570 Batswana family units with children. It measured computer and Internet use of each family member across 4 weeks. Data on actual computer and Internet usage were collected with the help of local leaders and teachers. They also played a key role in providing information on a number of variables for several age groups separately, including children, adolescents, and adult men and women. Averages were revealed for the amount of time spent on computers and the Web, the percentage of each age group online, and the types of Web sites viewed. Overall, about 9% of children ages 4 to 12, 40% of adolescents, 45% of adult women and 70% of adult men access the Internet each week. Children spend an average of 9 hours/week on the computer, 38 hours/week for adolescents. Adult women (non- working) spend only about 2 hours per week, yet in general, women were found to be spending long hours (25 hours) on computers corresponding with adult men who also spend 25 hours/week. The types of Web sites visited are accounted, including the top five for each age group. In general, search engines and Web portals are the most commonly visited sites, regardless of age group. These data provide a baseline for comparisons across time and cultures.***

Index Terms - Computer, Information Technology; Internet, Web, Search Engines

1. INTRODUCTION

Botswana has been among the best-performing economies over the past 35 years. During this period, Botswana has evolved from one of the poorest countries in the world to a middle-income country with the highest sovereign credit rating in Africa.

Botswana's Population (2005) was 1,765,000 with a literacy rate of 78.9% (2000-2004). According to a survey (2004), there connections out of 80,000 (75%) as against countries with very high user rates e.g. Norway (88%), Netherlands (87.8%), U.S.A (69.7%), Japan (67.1%), 57.7% (EU). Business leaders and scholars alike predicted that computer literacy will be as important in the 21st century as reading, writing, and arithmetic were in the 20th century (Anderson, Bikson, Law, & Mitchell, 1995; Peterson, 1995). As Slater (1994) among others suggests, information "will be the new wealth of the 21st century" (p. 96).

It is impossible to deny the tremendous effect rapid technological growth has had on our society. This explosion of new technologies has changed the way we live-from the way we do business and, to the way we communicate with each other. Technological advancements are also affecting the way we teach and learn. The business world demands that our schools prepare educated workers who can use technology effectively in the global market. (Technology and the New Professional Teacher, 1997)

Technology has undoubtedly entered the houses of Batswana families through the education system and government/non government offices. Hence people find themselves in a position where they do not have much choice but to learn it. The computer and internet have become so vital to business, education, communication, and entertainment that they no longer can be viewed as a luxury of the few and the rich; they are a crucial resource for all. The consequences of the use of computer and Internet on society are multifaceted and long term, and as a result, appreciating how families use them is important. The present study provides baseline information on how Batswana families use computers at home. Technology controls behaviour; even the simplest form of technology can influence individuals by restricting certain behaviours through enforcing rules. Latour's description illustrated how technology is engineered to limit or impose behaviours. However, much remains to be learned about "digital behaviours." The impact of Information Technology (IT) on individuals and society has only recently begun to be studied systematically. For example, there is a growing body of research on

1. the effects of violent video games on aggressive thoughts, feelings, and behaviours.
2. the effects of video games on visual attention
3. video game play and surgical skills
4. the uses of information technology to create digital archives of health and social information
5. the potential for pathological use of the Internet, computers, and video games.

These and other demonstrated potential effects of digital media demonstrate the importance of having some statistical data about the distribution and use of computer and Internet inside individuals' homes in Botswana. Parents are often the ones who pay for computers and Internet connections, and usually they are the ones who decide where to place them in the home. Interestingly, it has been found parents place the computer either in family rooms or in children's bedrooms. In addition, parents rarely set boundaries for the use of video games and

[1]Senior Lecturer, Tonota College of Education, Botswana
[2]Department of Mechanical Engineering, Faculty of Engineering & Technology, University of Botswana
E-mail: [1]srivastavarama@hotmail.com

computer/Internet. Despite the stereotypical view of older adults as resistant to new technologies, a growing number of national surveys report that the number of elderly who use computers has greatly increased since the early 1990s (Adler, 1996; Furlong, 1989; Hendrix, 2000; Lustbader, 1997; McKeely, 1991; Post, 1996; ScniorNet, 1998a, 1998b).

There is a favourable policy environment in Botswana. There has been a high level of interest in small, medium and micro-enterprise (SMME) development in Botswana since the approval of a new SMME policy in December 1998 (Ministry of Commerce and Industry, 1999). The policy's prime aim is to encourage further expansion of the SMME sector. In addition, Botswana's Vision 2016 document (Presidential Task Group, 1997) makes a strong commitment to the development of competitive enterprises through use of modern technology, including the implementation of IT across all manufacturing and service sectors. Similarly in schools also a number of computers have been installed for the use of students. The 'computerized surrounding' at work place and schools also builds the right environment to have this modern machine at home. Botswana had a tele density of 4.83 main lines per 100 inhabitants, compared to a low-income country average of 2.48. It had 33.42 Internet users per 10,000 inhabitants, compared to a low-income country average of 0.89 (ITU, 1998).

Richard Kassissieh, Director of Information Services at Catlin Gabel School in Portland, Oregon, U.S.A. reported in his report on Maru-a-Pula survey statistics that 81% of Form 1 students have a computer at home. 15% have broadband Internet access, 55% dial-up, and 30% no Internet access. These figures are in sharp contrast to American private schools, in which all students have a home computer connected to the Internet, and the majority has high-speed access. 75% of Form 1 students share their home computer with three or more users. Yet 49% still complete most of their computer-based schoolwork at home.

2. METHOD

2.1 Participants

Data were collected April through May 2008. The sample was a representative panel of Batswana families, including individuals from age 4 to over 60. The sample size was 570 individuals of whom 318 were male and 252 were female. In each district about 0.03% of the total population was selected to be part of the panel. Data were collected across a 4 week period for enhanced generalizability. The sample was initially contacted by random-digit dialling, which generated an equal probability sample of residential phone numbers. Families were interviewed during the initial enlistment call to identify households eligible to be in the panel (i.e., they have a computer and Internet access at home). At least 7 contact attempts were made to ensure that all eligible families were identified and recruited. Families that agreed to participate in the panel were mailed a documentation packet including the questionnaire and instructions.

2.2 Measurement of Independent Variables

Indexes were developed to measure certain independent variables like Age, Family Education Status and Monthly Income.

2.3 Measurement of Internet Activity

Measurement of Internet activity was designed to be in form of a 'schedule'. After completion of a demographic profile, panel member families with multiple computer users were asked to enter the details of particular active member and time of using computer/internet in two separate sections. All data were recorded by the active members who gave their detailed schedules of their computer and internet use. The data reported here were completely based on self-reports on actual measurements of computer activity by each participating person.

2.4 Population Reporting Method

The data were generalized to the general Batswana population from the panel. To provide population statistics, each panellist's data were categorised on the basis of gender, age, household income, education level, and region of the country. Data were grouped in six categories: children aged 4–11, adolescents 12–20, males 21–44, males 45+, females 21–44, and females 45+.

3. RESULTS

3.1 Demographic Details

Data presented in table 3.1 depicts that more than one third of the sample were adolescents (38%) and 56 per cent were males. Majority of the families had high educational status and monthly income of more than P7, 000. Majority of them were from Central, South east and Southern regions as these were the most densely populated regions of Botswana.

3.2 Average Weekly Time Spent on Computer and Internet

Table 3.2 illustrates the average hours spent on computers and the Internet for the total population and each age group. Surprisingly, even young children accessed the computer on average thrice in two days for a total of 9 hours per week, and adolescents' averaged more than five sessions in two days for an average of 38 hours per week spent on the computer. Adolescents' computer use was found to be much higher than the two groups of adult males. Adolescents, men and working women were found to be heavier computer users than the non-working two days for a total of 9 hours per week, and adolescents' averaged more than five sessions in two days for an average of 38 hours per week spent on the computer. Adolescents' computer use was found to be much higher than the two groups of adult males. Adolescents, men and working women were found to be heavier computer users than the non-working (housewives) adult females. This pattern is comparable when examining the amount of computer time specifically spent on the Internet. Young children spend an average of 32 hours per month on the Internet, whilst adolescents almost quadrupled that time (128 h/mo). It was interesting that use of the Internet as a percentage of total computer time increased as the total amount of computer time increased. It was heartening to note that the average use of home computer by users in Botswana was 24.5 hr/week.

3.3Average Time Spent on Web Site Categories

The Web sites visited were categorized as by Nielsen into 15 types. Table 3.3 displays the percentages of total time spent by each age group and gender (averaged across April-May 2008).It showed a pattern of adolescents aged 4 to 12 spending maximum time for entertainment (42%) whereas adolescents gave one fourth of their total computer time on Telecom and Internet services. Web pages related to search Engines, Entertainment and Telecom/. Internet services in general were found to be most popular amongst all the users (22%, 17% and 25% respectively). It was interesting to note that some websites were seldom browsed e.g. Multi-category commerce (2%) Government & non-profit (2%), Special occasions (2%), Travel (3%), Finance/ insurance/ investment (2%), corporate information (2%), Home & fashion (3%), Automotive (3%). Websites related to Education & careers (6%), News & information (5%), Family & lifestyles (5%) and Computers/consumer electronics (4%) were not very popular amongst the internet users.

Women were found to be the heaviest users (26%) of Telecom/Internet services beaten only by men of the same age group (27.5%). Search engines/portals, Entertainment and Telecom/Internet services constituted about thirty nine percent of the total time spent on Internet. Rest of the categories were browsed rarely by all the age groups.

4. DISCUSSION

The purpose of this study was to explore how Batswana families use the Internet from home. The study revealed that female were the heaviest users (26%) of Telecom/Internet services comparable with males of the same age group (27.5%). Search engines/portals, Entertainment and Telecom/Internet services constituted about thirty nine percent of the total time spent on Internet. Rest of the categories were browsed rarely by all the age groups.

Point to be noted is that children aged 4 to 12 spend 89% of computer time on the internet. An average child was found to be spending 11 hours a week on the computer, with an average of 7 Web sessions totalling 8 hours online. Children, like all age groups included in this study, were found to be accessing search engines and Web portals most frequently, but they were also accessing game and entertainment sites at a high rate. Adolescents aged 13 to 20 on the other hand accessed the Internet for 84% of net computing time. It was also evident that an average adolescent was spending 38 hours a week on the computer, with an average of 15 Web sessions totalling 32 hours online. Adolescents were accessing informational Web sites at a much higher rate than younger children, including a higher rate of accessing commercial Web sites each month.

Data were collected on two age groups of adult men in this study: 21 to 44 and 45 above. The two groups were quite comparable, with almost 8 out of 10 (79% and 86%, respectively) spending their computing time on the Internet. An average adult male was spending over 25 hours a week on the computer (29 and 21) for the two age groups respectively, with an average of about 19 Web sessions per month totalling to approximately 20 hours (23 and 18 respectively). Adult men were following the trend seen in adolescents: a high percentage of men accessing informational Web sites each month.

S.N	Attributes	Categories	Frequency (%), N=570
1	Age	4-12	126 (22.10)
		13-20	217 (38.07)
		21-44	142 (24.91)
		45 and above	85 (14.91)
2	Gender	Male	318 (55.78)
		Female	252 (44.21)
3	Family Education Status	Low	28 (4.91)
		Medium	142 (24.91)
		High	399 (70.00)
4	Monthly Income	< P3,500	0 (0)
		P 3,501-P 7,000	0 (0)
		>P 7000	570(100.00)
5	Region of the Country	Central	169 (29.6)
		Ghanzi	11 (1.92)
		Kgalagadi	6 (1.05)
		Kgatleng	12 (2.10)
		Kweneng	45 (7.89)
		North-East	68 (11.93)
		North-West	56 (9.82)
		South-East	102 (17.89)
		Southern	101(17.71)

Table 3: 1 Demographic Details

Similarly, we gathered data on two age groups of adult women in this study: 21 to 44 and 45 above. Adult women tended to access the computer less than men, and although the two groups of women were similar, they do not appear to be as similar as the adult men. In general, the younger group was more likely to spend time on the computer and the Internet than the older group. Women spent about 25 hours a week on the computer (32 and 18 hours for the two age groups respectively), with about 20 Web sessions per month (25 and 15 respectively), totalling 20 and 14 hours online per month respectively. The most exciting fact emerging from the study was a meagre time spent on computers and internet by non working (housewives). This also indicated that working women were having very long hours spent on computers which helped to maintain the average number of hours on computers still so high.

Adult women accessed Web site categories in a pattern similar to adult men, although at a lower rate than men and with a smaller percentage of women 45 above accessing each category of Web site compared to women 21 to 44.When examining the percentage of people accessing the Internet by day, it was striking how little variation there is across days. This suggests that perhaps the Internet is so integrated into people's lives that it is a daily habit, which is also indicated by the average number of Internet sessions per week being at least 15 for people over 11 years old.

5. CONCLUSION

The study revealed that women were the heaviest users of Telecom/Internet services beaten only by men of the same age group. Search engines/portals, Entertainment and Telecom/Internet services amounted to about two-third of the total time spent on Internet. In a nutshell an average child was found to be spending 11 hours a week on the computer, with an average of 7 Web sessions totalling 8 hours online. Children, like all age groups included in this study, were found to be accessing search engines and Web portals most frequently, but they were also accessing game and entertainment sites at a high rate. Adolescents accessed the Internet for 84% of net computing time. It was also apparent that an average adolescent was spending 38 hours a week on the computer, with an average of 15 Web sessions totalling to 32 hours online.

The two groups of adult men displayed quite a similar pattern with almost 8 out of 10, spending their computing time on the Internet. An average adult male was spending over 25 hours a week on the computer, with an average of about 19 Web sessions per month totalling to approximately 20 hours. Adult men were following the trend seen in adolescents: a high percentage of men accessing informational Web sites each month. Correspondingly, two age groups of adult women in this study tended to access the computer less than men, and although the two groups of women were similar, they do not appear to be as similar as the adult men. In general, the younger group was more likely to spend time on the computer and the Internet than the older group. Women spent about 25 hours a week on the computer with about 20 Web sessions per month. The most exciting fact emerging from the study was a meagre time spent on computers and internet by non working (housewives) which pointed to the fact that working women were having very long hours spent on computers which helped to maintain the average number of hours on computers still so high.

REFERENCES

[1] Adler, R. P. (1996). Older adults and computers: Report o a national survey. Retrieved November 17, 2004, from http://www.seniornet.org.

[2] Anderson, R. H., Bikson, T. K., Law, S. A., & Mitchell, B. M. (1995). Universal access to e-mail: Feasibility and societal implications. Santa Monica, CA: Rand Corporation

[3] Furlong, M. (1989). An electronic community for older adults: The Senior Net network. Journal of Communication, 39, 145-153.

[4] Hendrix, C. C. (2000). Computer use among elderly people. Computers in Nursing, 18, 1-13.

[5] http://www.bankofbotswana.bw/ Public Information Notice (PIN) No.04/78 July 30, 2004

[6] http://www.blogo.it/post/rilevazioniaudiweb- december-2005

[7] http://www. kassblog.com/ item / 250

[8] Latour B. (1992) Where are the missing masses? The sociology of a few mundane artifacts. In W.E. Bijker & J. Law (Eds.), Shaping Technology/Building Society: Studies in sociotechnical change. Cambridge, MA: MIT Press, p. 225.

[9] Lustbader, W. (1997). On bringing older people into the computer age. Generations: The Journal of the Western Gerontological Society, 21, 30-31.

[10] McNeely, E. (1991). Computer-assisted instruction and the older-adult learner. Educational Gerontology, 17, 229-237.

[11] Post, J. A. (1996). Internet resources on aging: Seniors on the net. The Gerontologist, 36, 565-569.

[12] SeniorNet. (1998a). SeniorNet Survey on Internet Use, November 2002. Retrieved from http://www.seniornet.org/php/default.php.

[13] SeniorNet. (1998b). Research on senior's computer and Internet usage: Report of a national survey. Retrieved from http://www.seniornet.org/php/default.php.

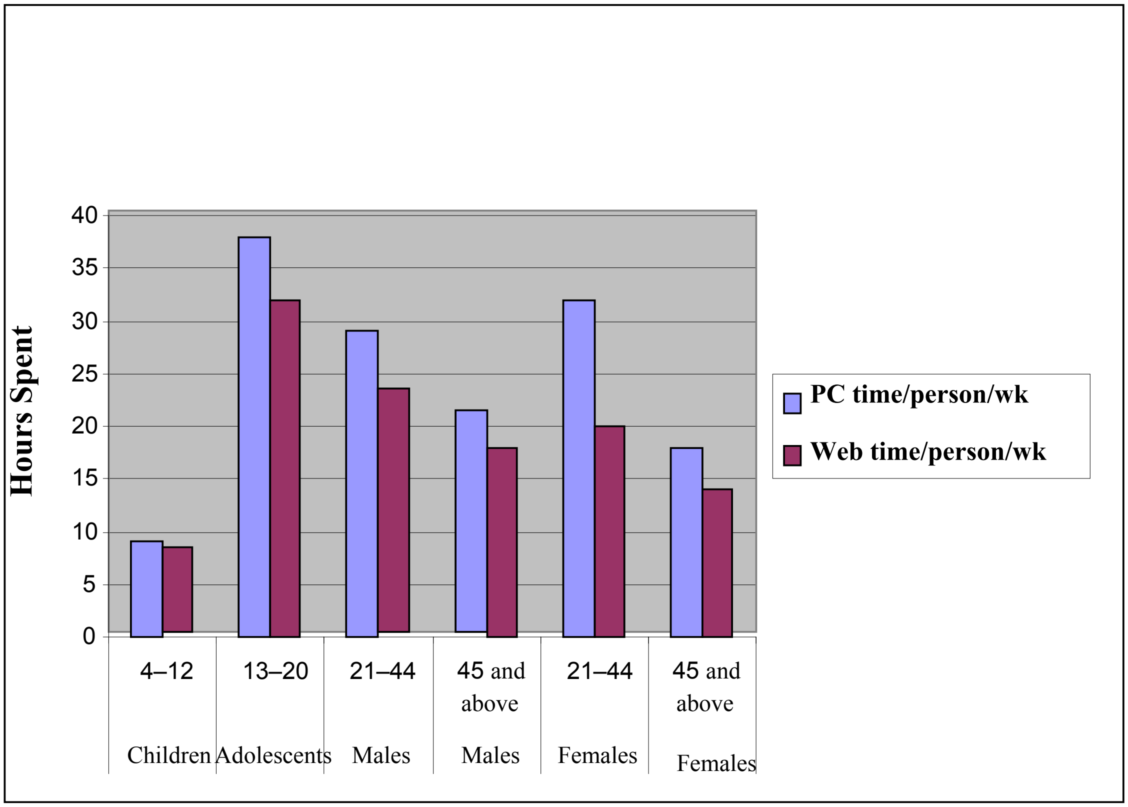


Figure 3.1: Time spent on Computers and Internet

	Children 4–12	Adolescents 13–20	Males 21–44	Males 45 and above	Females 21–44	Females 45 and above	Average
PC sessions/person/wk	11	18	23	21	28	18	19.83
PC time/person/wk(Hrs)	9	38	29	21	32	18	24.50
Web sessions/person/wk	7	15	19	18	25	15	16.50
Web time/person/wk(Hrs)	8	32	23	18	20	14	19.17
Web pages/person/wk	102	826	375	296	520	231	391.67
Percent Web time	89	84	79	86	63	78	79.83

Table 3.2: Average Use of Home Computers and the Web use, Split by Groups

Continued from page no. 78

[10] Margaret H. Dunham, Vijay Kumar (1998), Location Dependent Data and its Management in Mobile Databases, In. Proceedings of DEXA Workshop, Vienna, Austria, August 1998, pp. 414-419.

[11] H. Wachter and A. Reuter (1992), the ConTract Model, in [14], Chapter-7, pp. 219-263.

[12] J. Klein and A. Reuter (1988), Migrating transactions, in Proceedings of IEEE Workshop on the failure trends of Distributed Computing Systems, Hong Kong, Sept.

[13] F.M. Assis Silva (1999), A transaction Model based on Mobile Agents, PhD Thesis, Technische Universitat Berlin, Germany.

[14] A.K. Elmagarmid, ed. (1992), Database Transaction Models for Advance Applications, Morgan-Kaufmann Publishers, USA.

[15] R. B. Patel, V. K. Katiayar, Vishal Garg (2005), Mobile Agents in Wireless LAN and Cellular Data Networks, Journal of Computer Science, Science Publications New York, USA, 2(2).

[16] R. Ben-Natan (1995), CORBA - A Guide to Common Object Request Broker Architecture, McGraw-Hill.

[17] R. B. Patel and Neetu Aggarwal (2006), Load Balancing on Open Networks: A Mobile Agent Approach, Journal of Computer Science, Science Publications New York, USA, 2(5): 410-418, 2006.

[18] F.M. Assis Silva and S. Krause (1997), A Distributed Transaction Model based on Mobile Agents, in Proceedings of First International Workshop, MA'97, Berlin, Germany, April, Lecture notes in Computer Science 1219, eds. K. Rothermel, R. Popescu-Zeletin, Springer-Verlang, pp.198-209.

[19] J. Klein and A. Reuter (1988), Migrating transactions, in Proceedings of IEEE Workshop on the failure trends of Distributed Computing Systems, Hong Kong, Sept.

[20] R. B. Patel and K. Garg (2001), PMADE – A Platform for Mobile Agent Distribution & Execution, in Proceedings of 5th World MultiConference on Systemics, Cybernetics and Informatics (SCI2001) and 7th International Conference on Information System Analysis and Synthesis (ISAS 2001), Orlando, Florida, USA, July 22-25, 2001, Vol. IV, pp. 287-292.

[21] Moss, J. E. B (1981) Nested transactions: An approach to reliable distributed computing.PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA, April.

[22] Pu, C., Kaiser,G., and Hutchinson, N. (1988), Split Transaction for Open-ended Activities, in proceedings of the fourteenth International Conference on very large database, pp. 26-37, Sept.

[23] Buchman, A. et al. (1992) A Transaction Model for active Distributed Object Systems. In Elmagarmid, A. K. (Ed), Database Transaction Models for Advanced Applications, pp. 123-158. Morgan Kaufmann.

[24] Garcia-Molina, H. and Salem, K. SAGAS (1987), in Proceedings of the ACM SIGMOD International Conference on Management of Data, PP. 249-259, May.

[25] D.B. Terry (1985), Distributed Name Servers: Naming and Caching in Large Distributed Computing Environments, Ph.D. thesis, University of California, Berkely, 1985. Available as UCB/CSD Tech. Rep 85-228 and as Xerox PARC Tech. Rep. CSL-85-1.

[26] D. Barbara and T. Imielinski (1994), Sleepers and Workaholics: Caching Strategies in Mobile Environments, in Proceedings of ACM SIGMOD Conference Management of Data, pp. 1-12, May.

[27] R. Elmasri, S.B. Navathe (2000), Fundamentals of Database Systems , 3rd edition, 2000, Pearson Education, pp. 661-686, 694-700.

[28] S. Frenz, M. Schoettner, R. Goeckelmann, P. Schulthess (2004), Performance Evaluation of Transactional DSM, in Proceedings of the 4th IEEE/ACM International Symposium on Cluster Computing and the Grid, Chicago.

[29] M.Wende et al. (2002), Optimistic Synchronization and Transactional Consistency in Proceedings of the 4th International. Workshop on Software Distributed Shared Memory, Berlin.

BIJIT - BVICAM's International Journal of Information Technology

Paper Structure and Formatting Guidelines for Authors

BIJIT is a peer reviewed refereed bi-annual research journal having **ISSN 0973-5658**, being published since 2009, in both, Hard Copy as well as Soft copy. Two issues; **January – June** and **July – December**, are published every year. The journal intends to disseminate original scientific research and knowledge in the field of, primarily, Computer Science and Information Technology and, generally, all interdisciplinary streams of Engineering Sciences. **Original** and **unpublished** research papers, based on theoretical or experimental works, are published in BIJIT. We publish two types of issues; **Regular Issues** and **Theme Based Special Issues.** Announcement regarding special issues is made from time to time, and once an issue is announced to be a Theme Based Special Issue, Regular Issue for that period will not be published.

Papers for Regular Issues of BIJIT can be submitted, round the year. After the detailed review process, when a paper is finally accepted, the decision regarding the issue in which the paper will be published, will be taken by the Editorial Board; and the author will be intimated accordingly. However, for Theme Based Special Issues, **time bound Special Call for Papers** will be announced and the same will be applicable for that specific issue only.

Submission of a paper implies that the work described has not been published previously (except in the form of an abstract or academic thesis) and is not under consideration for publication elsewhere. The submission should be approved by all the authors of the paper. If a paper is finally accepted, the authorities, where the work had been carried out, shall be responsible for not publishing the work elsewhere in the same form. *Paper, once submitted for consideration in BIJIT, cannot be withdrawn unless the same is finally rejected.*

1. Paper Submission

Authors will be required to submit, MS-Word compatible (.doc, .docx), papers electronically *after logging in at our portal and accessing the submit paper link*, available at http://www.bvicam.ac.in/bijit/SubmitPaper.asp. Once the paper is uploaded successfully, our automated Paper Submission System assigns a Unique Paper ID, acknowledges it on the screen and also sends an acknowledgement email to the author at her / his registered email ID. Consequent upon this, the authors can check the status of their papers at the portal itself, in the Member Area, after login, and can also submit revised paper, based on the review remarks, from member area itself. The authors must quote / refer the paper ID in all future correspondences. Kindly note that we do not accept E-Mailic submission. To understand the detailed step by step procedure for submitting a paper, click at http://www.bvicam.ac.in/BIJIT/guidelines.asp.

2. Paper Structure and Format

While preparing and formatting papers, authors must confirm to the under-mentioned MS-Word (.doc, .docx) format:-

- The total length of the paper, including references and appendices, must not exceed **six (06) Letter Size pages**. It should be typed on one-side with double column, single-line spacing, 10 font size, Times New Roman, in MS Word.
- The Top Margin should be 1”, Bottom 1”, Left 0.6”, and Right 0.6”. Page layout should be portrait with 0.5 Header and Footer margins. Select the option for different Headers and Footers for Odd and Even pages and different for First page in Layout (under Page Setup menu option of MS Word). Authors are not supposed to write anything in the footer.
- The title should appear in single column on the first page in 14 Font size, below which the name of the author(s), in bold, should be provided centrally aligned in 12 font size. The affiliations of all the authors and their E-mail IDs should be provided in the footer section of the first column, as shown in the template.
- To avoid unnecessary errors, the authors are strongly advised to use the "spell-check" and "grammar-check" functions of the word processor.
- The complete template has been prepared, which can be used for paper structuring and formatting, and is available at http://www.bvicam.ac.in/BIJIT/Downloads/Template_For_Full_Paper_BIJIT.pdf.
- The structure of the paper should be based on the following details:-

Essential Title Page Information

- **Title:** Title should be Concise and informative. Avoid abbreviations and formulae to the extent possible.
- **Authors’ Names** and **Affiliations:** Present the authors' affiliation addresses (where the actual work was done) in the footer section of the first column. Indicate all affiliations with a lower-case superscript letter immediately after the author's name

and in front of the appropriate address. Provide the full postal address of each affiliation, including the country name and e-mail address of each author.
- **Corresponding Author:** Clearly indicate who will handle correspondence at all stages of refereeing and publication. Ensure that phone numbers (with country and area code) are provided, in addition to the e-mail address and the complete postal address.

Abstract

A concise abstract not exceeding 200 words is required. The abstract should state briefly the purpose of the research, the principal results and major conclusions. References and non-standard or uncommon abbreviations should be avoided. As a last paragraph of the abstract, 05 to 10 Index Terms, in alphabetic order, under the heading Index Terms ***(Index Terms - …….)*** must be provided.

NOMENCLATURE

Define all the abbreviations that are used in the paper and present a list of abbreviations with their definition in Nomenclature section. Ensure consistency of abbreviations throughout the article. Do not use any abbreviation in the paper, which has not been defined and listed in Nomenclature section.

Subdivision - numbered sections

Divide paper into numbered Sections as 1, 2, 3, …... and its heading should be written in CAPITAL LETTERS, bold faced. The subsections should be numbered as 1.1 (then 1.1.1, 1.1.2, ...), 1.2, etc. and its heading should be written in Title Case, bold faced and should appear in separate line. The Abstract, Nomenclature, Appendix, Acknowledgement and References will not be included in section numbering. In fact, section numbering will start from Introduction and will continue till Conclusion. All headings of sections and subsections should be left aligned.

INTRODUCTION

State the objectives of the work and provide an adequate background, with a detailed literature survey or a summary of the results.

Theory/Calculation

A Theory Section should extend, not repeat the information discussed in Introduction. In contrast, a Calculation Section represents a practical development from a theoretical basis.

RESULT

Results should be clear and concise.

DISCUSSION

This section should explore the importance of the results of the work, not repeat them. A combined Results and Discussion section is often appropriate.

CONCLUSION AND FUTURE SCOPE

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

APPENDIX

If there are multiple appendices, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similar nomenclature should be followed for tables and figures: Table A.1; Fig. A.1, etc.

ACKNOWLEDGEMENT

If desired, authors may provide acknowledgements at the end of the article, before the references. The organizations / individuals who provided help during the research (e.g. providing language help, writing assistance, proof reading the article, sponsoring the research, etc.) may be acknowledged here.

REFERENCES

Citation in text

Please ensure that every reference cited in the text is also present in the reference list (and vice versa). The references in the reference list should follow the standard IEEE reference style of the journal and citation of a reference.

Web references

As a minimum, the full URL should be given and the date when the reference was last accessed. Any further information, if known (DOI, author names, dates, reference to a source publication, etc.), should also be given. Web references can be listed separately (e.g., after the reference list) under a different heading if desired, or can be included in the reference list, as well.

Reference style

Text: Indicate references by number(s) in square brackets in line with the text. The actual authors can be referred to, but the reference number(s) must always be given. Example: '..... as demonstrated [3,6]. Barnaby and Jones [8] obtained a different result ....'

List: Number the references (numbers in square brackets) in the list, according to the order in which they appear in the text.

Two sample examples, for writing reference list, are given hereunder:-

Reference to a journal publication:

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "Secure spread-spectrum watermarking for multimedia", *IEEE Transactions on Image Processing*, Vol. 6, No. 12, pp. 64 – 69, December 1997.

Reference to a book:

[2] J. G. Proakis and D. G. Manolakis – Digital Signal Processing – Principles, Algorithms and Applications; Third Edition; Prentice Hall of India, 2003.

Mathematical Formulae

Present formulae using Equation editor in the line of normal text. Number consecutively any equations that have to be referred in the text

Captions and Numbering for Figure and Tables

Ensure that each figure / table has been numbered and captioned. Supply captions separately, *not attached to the figure*. A caption should comprise a brief title and a description of the illustration. Figures and tables should be numbered separately, but consecutively in accordance with their appearance in the text.

3. **Style for Illustrations**

All line drawings, images, photos, figures, etc. will be published in black and white, in Hard Copy of BIJIT. Authors will need to ensure that the letters, lines, etc. will remain legible, even after reducing the line drawings, images, photos, figures, etc. to a two-column width, as much as 4:1 from the original. However, in Soft Copy of the journal, line drawings, images, photos, figures, etc. may be published in colour, if requested. For this, authors will need to submit two types of Camera Ready Copy (CRC), after final acceptance of their paper, one for Hard Copy (compatible to black and white printing) and another for Soft Copy (compatible to colour printing).

4. **Referees**

Please submit, with the paper, the names, addresses, contact numbers and e-mail addresses of three potential referees. Note that the editor has sole right to decide whether or not the suggested reviewers are to be used.

5. **Copy Right**

Copyright of all accepted papers will belong to BIJIT and the author(s) must affirm that accepted Papers for publication in **BIJIT** must not be re-published elsewhere without the written consent of the editor. To comply with this policy, authors will be required to submit a signed copy of Copyright Transfer Form, available at http://bvicam.ac.in/bijit/Downloads/BIJIT-Copyright-Agreement.pdf, after acceptance of their paper, before the same is published.

6. **Final Proof of the Paper**
   One set of page proofs (as PDF files) will be sent by e-mail to the corresponding author or a link will be provided in the e-mail so that the authors can download the files themselves. These PDF proofs can be annotated; for this you need to download Adobe Reader version 7 (or higher) available free from http://get.adobe.com/reader. If authors do not wish to use the PDF annotations function, they may list the corrections and return them to BIJIT in an e-mail. Please list corrections quoting line number. If, for any reason, this is not possible, then mark the corrections and any other comments on a printout of the proof and then scan the pages having corrections and e-mail them back, within 05 days. Please use this proof only for checking the typesetting, editing, completeness and correctness of the text, tables and figures. Significant changes to the paper that has been accepted for publication will not be considered at this stage without prior permission. It is important to ensure that all corrections are sent back to us in one communication: please check carefully before replying, as inclusion of any subsequent corrections cannot be guaranteed. Proofreading is solely authors' responsibility. Note that BIJIT will proceed with the publication of paper, if no response is received within 05 days.

BVICAM'S International Journal of Information Technology

(A Biannual Publication)

Subscription Rates

	I Year		3 Years	
Category	**India**	**Abroad**	**India**	**Abroad**
Students	Rs. 400	US $ 45	Rs. 1000	US $ 120
Individuals	Rs. 300	US $ 40	Rs. 750	US $ 100
Institution	Rs. 250	US $ 30	Rs. 600	US $ 075
Companies	Rs. 150	US $ 025	Rs. 375	US $ 050
Single Copy	Rs. 250	US$ 025	-	-

Subscription Order Form

Please find attached herewith Demand Draft No._______________ dated __________________

For Rs.______________ drawn on ____Bank in favor of **Director, "Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi"** for a period of 01 year / 03 Years

Subscription Details

Name and Designation ____

Organization ____

Mailing Address ____

_______________________________________ PIN/ZIP ____________________________

Phone (with STD/ISD Code)_____________________________FAX___________________

E-Mail (in Capital Letters)____

Date: **Signature**

Place: (with official seal)

Filled in Subscription Order Form along with the required Demand Draft should be sent to the following address:-

Prof. M. N. Hoda
Director
Bharati Vidyapeeth's
Institute of Computer Applications & Management
A-4, Paschim Vihar, Rohtak Road, New Delhi-110063 (INDIA).
Tel./ Fax: 91 – 11 – 25275055 E-Mail: bijit@bvicam.ac.in
Visit us at: www.bvicam.ac.in

About Bharati Vidyapeeth: Bharati Vidyapeeth, which is our parent body, was established on 10th May 1964 by Hon' ble **Dr. Patangrao Kadam** with a wider objective of ***"Transformation through dynamic education"***. Under the leadership of the Hon' ble Founder, Bharati Vidyapeeth has made astonishing strides in the field of education, during a short span of 44 years with a network of more than 187 institutions all over India. Acknowledging the excellence, the Ministry of HRD, Govt. of India, on the recommendation of UGC, New Delhi, has accorded the status of a Deemed to be University to eleven faculties of Bharati Vidyapeeth in April 1996. By now, Bharati Vidyapeeth University, Pune, has 33 Institutions / Colleges as its constituent units. Besides making contribution to the intellectual awakening, its activities have been geared to bring multidimensional progress and welfare of different section of population including women, tribal and rural people.

About BVICAM*: Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)*, New Delhi, was established by Bharati Vidyapeeth, Pune*,* in the year 2002. BVICAM is a reputed and most sought after Institute for MCA programme in north India. It is approved by All India Council for Technical Education, New Delhi, and is affiliated to Guru Gobind Singh Indraprastha University (GGSIPU), Kashmere Gate, Delhi. Presently, it runs 03 years **Master of Computer Applications (MCA)** Programme. BVICAM is centrally located at National Highway No. 10, Rohtak Road, A-4, Paschim Vihar, New Delhi, in its own state of the art sprawling campus.

It has state of the art infrastructural and instructional facilities, comparable to the best in the world. BVICAM also contributes effectively towards providing excellent opportunities for teaching and research activities by organizing several National Seminars, Conferences, Symposiums, Faculty Development Programmes, Workshops and publication of Research Journals. In this sequel, BVICAM feels proud to release the Inaugural Issue of BVICAM's International Journal of Information Technology (BIJIT).

National Conference On
Computing For Nation Development
Bharati Vidyapeeth Educational Complex, New Delhi
2nd National Conference On
Nation Development
(February, 2008)
om-2008
Conference On