Name of the Scholar: **Pramod Kumar Yadav**

Name of the Supervisor: **Prof. (Dr.) Syed Afzal Murtaza Rizvi**

Department: **Department of Computer Science, Jamia Millia Islamia, New Delhi**.

Title: **Efficient Query Optimization for Mining Distributed Data**

Keywords: **Iterative Improvement, Simulated Annealing, Two Phase Optimization, Query Optimization, Distributed Database**.

Abstract

The functional arrangements of various nodes/workstations, distributed over a computer network is called Distributed Database where as the technique of finding the optimal query processing method to answer a query is called Query Optimization. In Distributed Database, the user or nodes communicates with each other through networks. The exponential growth of internet and their complexities are due to the fact of fast incrementation of user or nodes or any network. This phenomena invites enormous challenges as open problems to explore solution in near future. There are various issues arises during evaluation of query cost, among which the processing cost and a transmission cost are the important. Attempts are made to develop several algorithms to find the best possible solution for a particular query; however all these algorithms have their certain limitations. The primary objective of distributed query plan is to generate the optimal query plan that reduces the quantity of data transfer amongst sites and also the distributed query response time.

In distributed database system the data is scattered over various multiple sites and a single relation may be present in more than one sites. In such case the cost of query processing is effected by various cost involved such as optimization cost, communication cost, local processing cost, query localization cost etc. The optimizer is mainly concern on the cost model, search space, and the search strategy. It primarily focuses on these three factors. Hence, to find the optimal cost for a particular query is emerging as an open challenge for many researchers. Therefore the cost-based query optimization technique has emerged as an important concept for dealing with the query optimization. However using the concept of fragmentation and replication, the database is physically distributed across various sites. Fragmentation plays a vital role in dividing each relation into different fragments. In fragmentations, the relation may be replicated for each fragment to various distributed

sites, since the same data may be accessed from applications that executes at a number of sites. The present study has implemented the concepts of iterative improvement algorithm and simulated annealing algorithm separately and the optimal query plans have been generated by applying the concept of iterative improvement algorithm and simulated annealing algorithm. The comparative studies of both results are done based on the number of optimal query plan generated and the average query processing cost, using same heuristic. Therefore in this process the study have propose that the simulated annealing algorithm produces better optimal query plan as compared to iterative improvement algorithm and the average query processing cost in simulated annealing is lower as compared to iterative improvement algorithm.

Randomized algorithms such as Simulated Annealing and Iterative Improvement are viable alternatives to exhaustive search. Attempts have been developed to adapt them to find the optimal query plan in a distributed database. The present study has tested them on large queries, concluding that in most cases Simulated Annealing identifies a lower cost access plan than Iterative Improvement.

In this thesis, attempts are made to develop and implement the concept of Two Phase Query Optimization algorithm for generating the optimal query plan, which is also known as hybrid approach which works in two phase: in first phase iterative improvement algorithm is applied followed by simulated annealing algorithm. The two phase optimization algorithm is one of the best known randomized algorithms. The Experimental results show that Two Phase Optimization outperforms the results of Simulated Annealing and Iterative Improvement Algorithms in terms of generating optimal query plan. The results of the algorithm are compared on the two factors i.e. no. of query plans generated and average query processing cost using the same heuristic by varying the no. of relation and the no. of distributed sites participating. The present study propose that the two phase query optimization algorithm produces the best optimal query plan as compared to the iterative improvement and simulated annealing algorithm. The average query processing cost is maximum for iterative improvement followed by two phase query optimization algorithm, where as it is minimum for the simulated annealing algorithm.

Hence, the present study concludes that the two phase optimization algorithm is the best algorithm for generating optimal query plan in distributed environment as compared to Iterative Improvement algorithm and Simulated Annealing algorithm.