Query Optimization in Distributed Databases

Vaishnavi[1] ,Vikram Khandelwal[2]

[1]*B.Tech Scholar, Poornima Group of Institutions, Jaipur*

[2]*Assistant Professor Department of Computer Science, Poornima Group of Institutions, Jaipur*

Abstract

Query optimization is a way of implementing the best plan for the query so that the performance of the query can be improved. In case of distributed database query optimization is much difficult as compared to centralized database. Queries are affected by many factors in case of distributed database such as the methods of insertion of data into the server and the time of transmission between the server.
In this paper, various optimization strategies have been analyzed and the study shows that the performance of distributed query can be improved with the help of hybrids of ACO.

Keywords: *Join Query Optimization, Ant Colony Optimization*

1. Introduction

Asthe high speed communication network are being launched, researches are also increasing for developing highly efficient techniques for processing complex queries in a cost efficient manner in distributed database. A collection of inter-related database which is distributed over a network so that, their must be logical enhancement in the computer performance, reliability, availability and modularity of the distributed database.

As the data in distributed database is present at multiple sites present at different geographical locations, query processing also includes transmitting data among different sites. The process of fetching data from different geographical locations is known as Distributed Query Processing.

Mainly there are three phases involved in distributed query processing:-

I. Local Processing Phase: - In this phase, data decomposition takes place i.e, the algebraic queries that are specified in global relations are transformed into fragments and are made available to different sites for the purpose of processing.

II. Reduction Phase:-Here the size of joins that are to be transmitted for accomplishing the join operation are minimized in a cost effective manner.

III. Final Processing and Assembling Phase:- Here the final output is generated by transmitting all the processed files to the assembly sites.

The ability of a query optimizer of deriving the most efficient query processing strategies critically determines the performance of a distributed query. So, query optimization is considered as one of the most important stage in the execution of distributed queries. The number of relations of the initial query and the set of rules required for generating query trees, typically determines the complexity of the optimization process.

As the query is entered by the user, it is transformed into relational algebra, then the optimizer searches for the most efficient query execution plan. As the number of relations required for processing the query increases, there will be an exponential increment in the number of alternative queries. For the purpose of generating the optimal query plans, the query optimizer will explore the large search space. An attempt has been made in the given research paper so as to analyze the various search strategies so that the optimal query execution plan can be obtained for solving a query in distributed database.

2. Literature Review

ShyamPadia[1] proposed the Ant Colony Optimization Algorithm for optimization of queries in Distributed Database. The realization of hybrids of Ant Colony Optimization Algorithms in the optimization process of Distributed Database is still a new field and various researches are still in progress for creating and implementing the hybrids of Ant Colony Optimization problems. According to the proposed research there is a huge opportunity for generating optimized solutions using hybrids of ACO in Distributed Database.

PawandeepKaur[2] proposed the join query optimization in Distributed Database. In this method firstly the data from server site will be transformed to the client site, then the data will be inserted into the client database before performing the join operation. By this method the insertion time of data will be reduced. So, this method is for improving the time efficiency of query optimization in Distributed Database.

Ms. PreetiTiwari[3] proposed the Hybrids of Ant Colony Optimization Algorithm for query optimization in Distributed Database. According to her given review and studies the performance of Distributed Query can be enhanced by combining Ant Colony Optimization with other optimization algorithm.

3. Algorithms

Join Query in Distributed Database is used for joining data from different sites[1]. Join Query Optimization in centralized database is much simple as compared to Distributed database. Various work on join query optimization in distributed database has been done for calculating the size of data on two different sites that are present on two different locations and then to tranferthe data of smaller size on another site and then performing join operation.

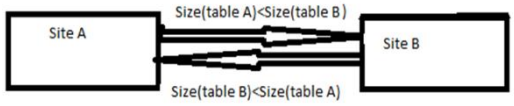

Fig : Mininized size of data transmission

Parallel Query Processing is another method for join query in distributed database[1]. The main focus of parallel processing is to maximize the number of simultaneous transmissions.

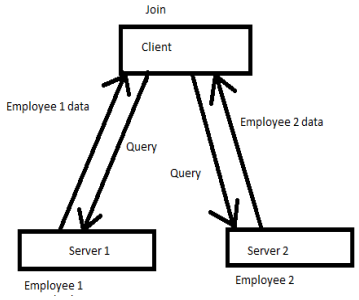

Fig 2: Parallel Processing of join query

Here, client send query for retrieving data from server1 and server2. After that server1 sends employee_1 data and server2 sends employee_2 data to the client. Then, after inserting data into database the client performs the join query on the data from two different servers. If server1 contains the employee_1 relation as:

Employee_1(employee_id, employee_name,employee_salary)

Employee_2(employee_id,Employee_address,employee_designation)

Now client wants to join the employee_1 and Emplouyee_2 relation from server1 and server2 respectively and also want to perform a query Q.

Q: SELECT* FROM EMPLOYEE_1,EMPLOYEE_2 WHERE S.EMPLOYEE_1id=EMPLOYEE_2id.

In distributed databases, query Q can be divided into three parts:

SELECT *FROM EMPLOYEE_1

SELECT*FROM EMPLOYEE_2

SELECT*FROM EMPLOYEE_1,EMPLOYEE_2

WHERE E1.id=E2.id

Here, the data from query1 and query2 are selected from two different tables. Since, the data resides on two different machines at different geographical location. So, the transmission of data is not required here, but Query3 is the join query that cannot be executed until the data from remote sites are transferred to the same sites.

In this case the QEPS are commonly known as join trees for which operators will be various types of joins, which will be represented as Query Graph and will be denoted as G = (N,A) when N will denote the set of nodes and A will denote the set of arcs. The nodes here will represent the set of Base Files in the join specifications of query. If a query joins the two corresponding nodes then this will be represented by an arc.

As for example let file F1 is stored at S1 and S2 sites, files F2 is stored at s3 site and file F3 is stored at site s4. Assume a query for joining the sequence of these files from F1 to F4. Join attributes will be given as: Select A1,<non-join projection attributes list> From F1,F2,F3,F4 where F1_A0=F2_A0 and F1_A1=F2_A1 and F3_A2=F4_A2.

The query graph for the given distributed query will be given as:

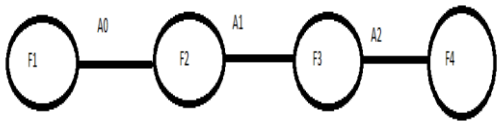

Fig 3: Query Graph for Distributed Query

Since, the given query graph have complex representation. It will be represented as follows:

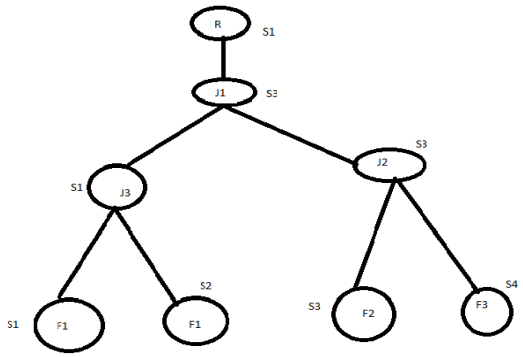

Fig 5(a): Optical

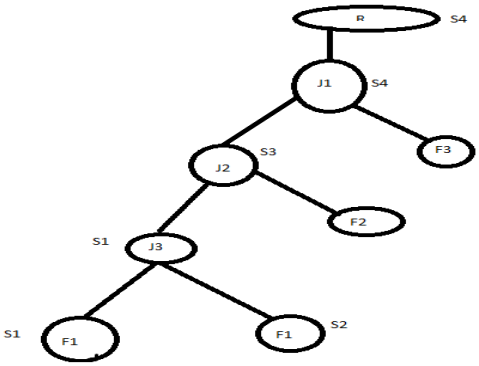

Fig 5(b): sequential level

Fig 5: Query Execution Plan for above Query

While performing the join query i.e., accessing data from different sites, the performance of the query must be enhanced by minimizing the cost. While fetching the result of query of data present at different sites, insertion and transmission cost are very important. Since, usually the insertion of data from server to client database takes more time than the transmission of data. So the main objective is to improve the performance of Join Query so that the insertion cost can be decreased.

Ant Colony Optimization Algorithm is a meta-heuristic algorithm which is good for the issues arising in combinatorial optimization[3]. Italian scholars Dorigo M, Colorni A and Maniezzo Vfirstly proposed Ant Colony Optimization Algorithm in 1992. ACO is a Bionic Algorithm which was influenced by Ants and are used for solving computational problems with the help of probabilistic technique. It is one of the most robust algorithm which provides intelligent searching and can also be used for Global Optimization solutions.

Though ACO possess special characteristics like positive feedback mechanism, robust nature and distributed computing, it has some deficiencies also:

1. There is no proposed systematic way to start the initial formation that is needed by ACO.
2. Because of the positive feedback mechanism the convergence speed of ACO increases towards best answer but it will be lower in the beginning because of the presence of small amount of pheromone.

A new Genetic Algorithm[3] based query optimization was proposed by the scientist. The newly proposed optimization was called as NGA as that improves the query execution process with the help of join orders, join sites and semi join reducers. In the given algorithm it was possible to reduce the local processing cost and Network Communication cost. Scientists also proposed a combinational algorithm of Genetic Algorithm and Learning Automata for producing the most suitable Query Execution plans based on Join Order Execution and Join Site Selection plans.

A combination algorithm of Genetic Algorithm and Heuristics was also proposed so as to solve the join order problem like Travelling Salesman Problem in large-scale database[3]. A Hybrid Algorithm was also proposed by using the properties of ACO & particle Swarn Optimization for solving the Travelling Salesman Problem. In this algorithm initially the statistics method was adopted for getting several better solutions initially. And then according to them information pheromone are distributed. Then with the use of ACO & information pheromone accumulation and renewal different solutions are obtained. Finally, the most effective solutions are obtained with the use of across and mutation operation of particle swarn optimization.

As the number of relations in a query increases, a huge memory and processor is also required. Now all the commercial applications which obtains data from different sites uses DDBMS as standard DBMS. For directing the ants towards the unexplored areas of search apace, the path which marks the behavior of ants are applied so as to visit all the nodes without knowing the graphic topology for generating the best solution for distributed database queries. Here, the ants will provide fast, high performance and best results in a cost effective manner by calculating the running time of the execution plans of the given query.

The Search strategy that are adopted by the Query Optimizer in the Distributed Database Management System will help it to increase the timing efficiency and cost efficiency and hence enhancing the performance of the query by choosing the best plan. With the use of these probabilistic algorithms the must viable solutions can be generated in case if the size of the query and the number of joins increases.

4. Conclusion

The realization of hybrids of Ant Colony Optimization Algorithm for the optimization of distributed database queries is still a very new field. The researches for

creating and implementing the hybrids of ACO are effective and viable in optimization problem. There is lots of opportunity for generating optimized solutions and for refining the search strategies with the use of hybrids of ACO for distributed database queries when the size of relations increases as per the increase in the number of parameters influencing the queries.

References

[1.] S Padia, and S Khulge(2015).Query Optimization Strategie in DistributedDatabases,5,4228-4234,India:Mumbai. International Journal of Computer Science and Information Technology.

[2.] P Kaur, and J K Sahiwal(2013 May). Join Query Optimization in Distributed Database,5,1-3,India:Phagwara. International Journal of Scientific and Research Publication.

[3.] P Tiwari, and S V. Chande(2013 June). Optimization of Distributed Database Queries Using Hybrids of Ant Colony Optimization Algorithm,3,609-614,India:Jaipur.International Journal of Advanced Research in Computer Science and Software Engineering.

[4.] K Kaur, and R K Sahu (2016 March). Query Optimization in Distributed Database,6,679-684,India:Punjab. International Journal of Advanced Research in Computer Science and Software Engineering.

[5.] R Sharma, P Verma, and S Chaudhary(2015 July). Query Optimization Concepts in Distributed Database,2,60-65,India:Muzaffarnagar. International Journal of Engineering Technology Science and Research