Advance Clustering Technique for Query Optimization in Distributed Database

Juhi Srivastava[1], Prof. Gayatri Pandi (Jain) [2].

[1]Student (Master of Engineering), Computer Engineering, L.J. Institute of Engineering and Technology, Gujarat, India.
[2] H.O.D.PG Departments, Computer Engineering, L.J. Institute of Engineering and Technology, Gujarat, India.

ABSTRACT

In modern days Distributed Database system suffers from many difficulties. One of the most common difficulties is the environment where such kinds of systems are running on unpredictable and volatile environment. In such an environment it is difficult to produce an efficient database query optimization based on information available at compilation time. In this paper a study of traditional genetic algorithm along with some new techniques are done for finding out the limitation and overcome them for optimizing query in Distributed Database. All the different work proposed in papers are based on genetic algorithm and new techniques are imposed on it to improve the performance of query execution. Some techniques involve clustering technique to minimize the cost for executing query and an ant system is used to improve the execution cost as ant takes optimal path to solve any problem. Finally new teacher- learner based techniques are used to optimize the query. In this paper we propose a model of query optimization which reduce time complexity by using GFCM (Geostatistical Fuzzy c-mean) in place of FCM (Fuzzy c-mean) clustering technique.

Keyword- *Database, Distributed Database, Query, Query Optimization, Genetic algorithm.*

1. INTRODUCTION

A Database is a collection of Schema, Relation or Data. It contains or holds the operational data that can be shared and accessed by concurrent user. [3] A Database can be Centralized, Relational, Object-Oriented, and Distributed etc. A traditional Distributed Database is defined as connection of several datasets that are scattered or dispersed physically but centralized logically with a combination of computer network and database systems.[2] This system are group of autonomous collaborating organizations that facilitates storing of information at physical distributed positions, depending on the frequency of admittance by consumers confined to a place. This collection of data is logically belonging to the same system but is distributed over different geographic sites of a computer network.[1]It encompasses coherent data spread across various sites of a computer network.[5] DDBMS(Distributed Database Management System) provides access to user via a simple and unified interface over disparate database, as if they were not distributed. [5]

The primary purpose of Distributed Database query optimization is to make the communication cost and response time of the query minimum that is to minimize the cost to obtain the required data in shortest possible time. [1] Query optimization is defined as a technique where the finest implementation approach of a specified query is obtained from a group of options, [2] and a Query is an enquiry into the database using Select statement. A Query is used to extract data from the database in a readable format according to the user's request.

2. RELATED WORKS

2.1 LITERATURE REVIEW

2.1.1Distributed Database query based on improved genetic algorithm

In [1] ShaoHua Liu, Xing Xu, the author uses FCM (Fuzzy c-mean) algorithm in addition to genetic algorithm to optimize the query of Distributed Database. In this algorithm it first creates a coding tree in which each non-leaf

node are represented as 0 and other leaf node with related table sequence number for a particular query. Then in second step it will apply evaluation function to calculate the corresponding cost of each chromosome through defined cost model. In third step, a new generation of individuals can be obtained which are combined with the characteristics of their parents. It reflects the idea of information exchange. Here all the contemporary individuals are divided into three categories through FCM clustering algorithm and each category is set to different crossover probability. Then in final step it randomly selects an individual in the group and change the value of one of the string structure data with a certain probability for the selected individual, this will provide the opportunity to generate new individual. Then it applies stopping condition, pre-given evolution algebra and chromosome string according to the fitness function is less than the given value which gives the optimal set of the problem.

2.1.2Query optimization using clustering and genetic algorithm for distributed database

In [2] S.Venkata Lakshmi, Dr. Valli Kumari Vatsavayi, two different phases are proposed in this paper. In first phase an evolutionary approach known as genetic algorithm is employed to obtain closely related or minimum number of different database sites for a given query, which is given as input to the second phase. In this phase a clustering technique is employed where the given query is matched with the existing database of query template cluster. Here author proposed this methodology for an efficient query optimization. The genetic algorithm is used to obtain the query plan containing the essential information that resides in lesser sites and leads to effective query processing. And in second phase, a clustering technique is employed which groups the related queries into clusters and employs the optimizer introduced strategy for the cluster demonstrative to implement whole upcoming query allocated to the cluster.

2.1.3Design and analysis of stochastic DSS query optimizer in a distributed database system.

In [3] Manik Sharma, Gurvinder Singh, Rajinder Singh, A distributed DSS query optimizer has been design to solve the operation site allocation problem of distributed DSS query. For finding an optimal operation site allocation plan, first of all a SQL based decision support system query is decomposed into relational algebra expression based on selection, projection, join and semi-join. In this paper, author uses stochastic approach and traditional genetic algorithm to generate new approach. It generate two approaches (i) Restricted Stochastic Query Optimizer (RSQO), randomly generate initial population, and then generates chromosomes to allocate sub-operation of a DSS query on a distributed network. This innovation lies in the restricted growth of chromosomes design and (ii) Entropy Based Restricted Stochastic Query Optimizer (ERSQO), uses Harvard and Charvat entropy to refrain low diversity population problem which normally occur in the implementation of Genetic algorithm.

2.1.4Query optimization of distributed database based on Parallel Genetic algorithm and Max-Min Ant system

In [4] Wenjiao Ban, Jiming Lin, Jichao Tong, Shiwen Li, firstly a set of optimal solution is produced by genetic algorithm in every processor and transform them into a certain amount of the initial pheromone. Then unify the initial solution of each ant colony. Finally execute MMAS (Max-Min Ant System) algorithm in parallel to get the more optimal solution. Here genetic algorithms fast convergence to take a set of relatively optimal QEPs (Query Execution Plans). Then MMAS guide ants to find the optimal QEP. Meanwhile process the hybrid algorithm in parallel to improve solving speed. PGA-MMAS full shows the superiority of parallelism, when the number of relationship is greater. The search time of optimal QEP of PGA-MMAS (Parallel Genetic Algorithm and Max-Min Ant System) relatively less compared with other algorithms and its high quality QEP also reduced the query execution time.

2.1.5 Generating Optimal Query plans for distributed Query processing using Teacher-Learner Based optimization

In [5] Vikash Mishra and Vikram Singh, the author uses parameter less optimization technique. This algorithm work for multi-objective unconstrained and constrained benchmark. This algorithm consider a group of learner as population and different subject offered to the learner are considered different design objective and a learner's result is analogous to the fitness value of the optimization problem. The best solution in the entire population is considered as the Teacher. TLBO start by initializing the entire set of query plan for given user or application query using pre-determined Relation Site Matrix, these query plans in solution space are equivalent to student or learners of TLBO then in teacher phase, student or learner learn via teacher, a teacher attempts to increase the mean result of the class in the subject depending on his or her capability and in student phase, each learner raise their level knowledge level by interaction among themselves. Then final selected QEP are used for result generation which is send back to the origin site from which user send the query and query plan is kept in directory for future reference.

2.2 COMPARATIVE TABLE

Table -1: Comparative Table

Sr. No.	Paper Title	Method Used	Advantages	Disadvantages
1	Distributed database query based on Improved Genetic Algorithm	FCM (Fuzzy c-mean) and traditional genetic algorithm	prevent the local optimum	Clustering algorithm used takes long computation time, sensitive to initial guess and sensitive to noise
2	Query optimization using Clustering and Genetic algorithm for distributed database	Clustering and genetic algorithm	Avoid redundant pre-processing of the plan space	Clustering of query based on similarity feature is time consuming
3	Design and Analysis of stochastic DSS query optimizers in a distributed database system	Decision support system queries optimizer(EAQQ, SGQQ, NGQQ, RSQO, ERSQQ)	Enhance the performance of traditional genetic algorithm	Quality of Service(Total cost) of SGQQ and NGQQ is not good
4	Query optimization of distributed database based on parallel Genetic algorithm and Max-Min ant system	Genetic algorithm and max-min ant system (parallel)	Solve the problem of easier to fall into an early-maturing state and local optimum reduced.	Parallel computation of both algorithm is required, increases the computation cost.
5	Generating optimal query plans for distributed query processing using Teacher-Learner Based optimization	Parameter less optimization technique	Effective and robust and has a great potential for solving similar multi-objective problems.	Does not work for similar design objectives.

3. CURRENT ISSUES IN SYSTEM

According to the literature review there are some problems related to Query optimization in Distributed Database such as: high Time complexity, need concern while Selection of clusters, low Quality of service, node configuration etc. The proposed model's main focus is to improve time complexity. That is main goal is to reduce query execution time over different sites.

4. PROPOSED WORK

The proposed Model shown in Fig.1 starts with taking client's query and sharing of client data to Distributed Server which can be in form of load or chunk distribution. Then the whole process is divided into two phase (i) clustering of data using GFCM (Geo-Statistical Fuzzy c-mean) and (ii) apply traditional genetic algorithm to solve the query optimization problem. GFCM gives centroid from several numbers of query plan and then traditional genetic algorithm is applied on the selected query plan to get faster result. In case it will not converge fast a Taylor series execution mechanism will going to be used to get faster result.

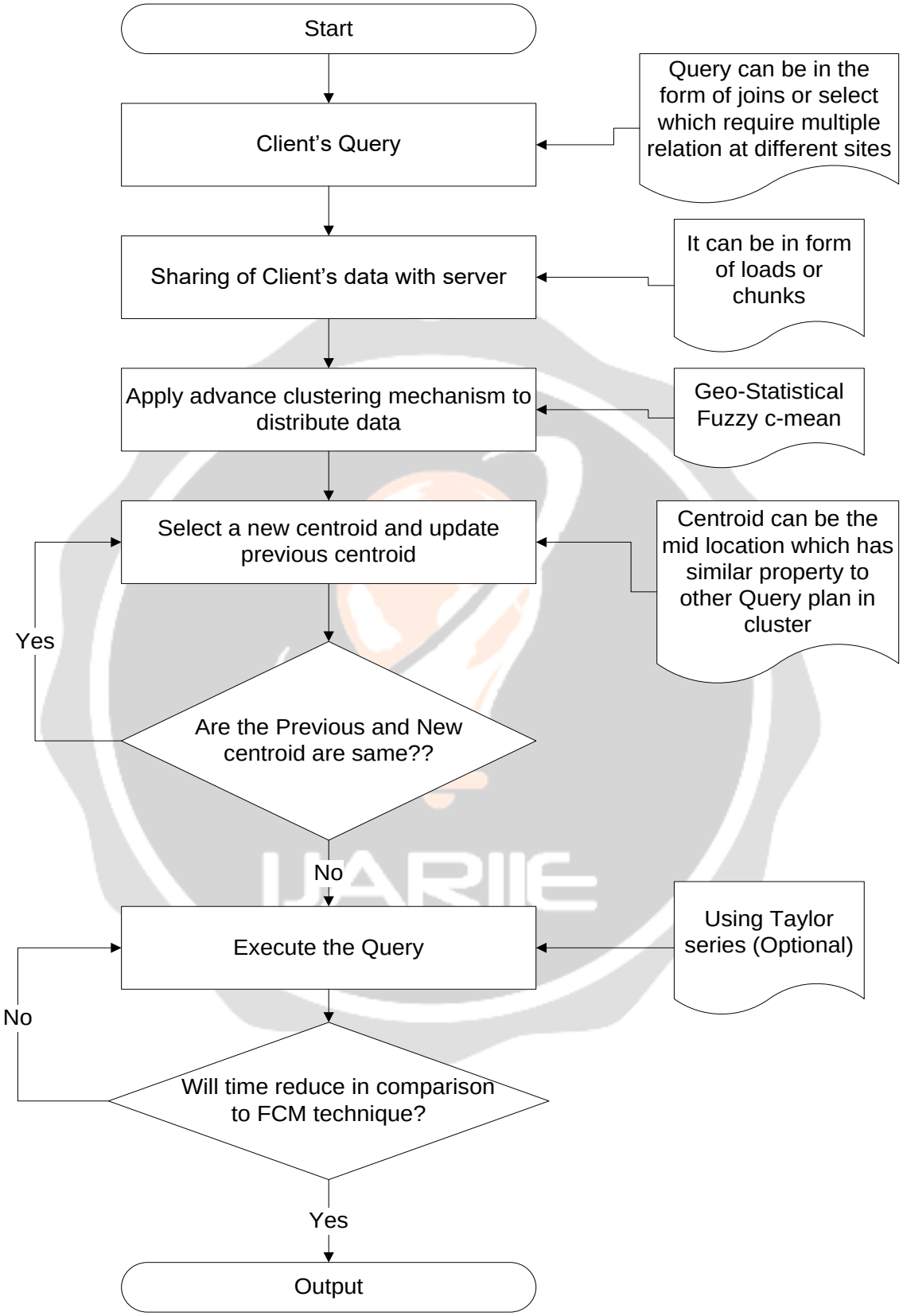

Fig - 1: Flow chart of Proposed Model

5. CONCLUSIONS

We go through a distributed database system in which data is dispersed over the multiple sites, this distribution of data is based on partition or replication based due to which a given relation can be found in more than one location. Query processing in such an environment require more CPU power, I/O and site- to- site communication cost, there will be need to reduce the time for query execution i.e. query optimization. There are certain research areas in such field like high execution time, QoS, clustering of data etc.

6. ACKNOWLEDGEMENTS

The authors are very much grateful to Department of Computer Engineering, L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; for giving opportunity to do research work on Query optimization in database. Two authors Juhi Srivastava, and Prof. Gayatri Pandi (Jain) are also grateful to management team of L.J Institute of Engineering & Technology, Ahmedabad, from Gujarat Technological University; for giving constant encouragement to do research work in the Department.

7. REFERENCES

[1] ShaoHua Liu, Xing Xu "Distributed Database Query Based on Improved Genetic Algorithm" International conference on Information Science and Control Engineering , pp. 348-351,2016.

[2] S.Venkata Lakshmi, Dr. Valli Kumari Vatsavayi "Query Optimization using Clustering and Genetic Algorithm for Distributed Databases" International Conference on Computer Communication and Informatics, 2016.

[3] Manik Sharma, Gurvinder Singh, Rajinder Singh "Design and analysis of stochastic DSS Query Optimizers in a distributed database system" Egyptian Informatics Journal, pp. 161-173, 2016

[4] Wenjiao Ban, Jiming Lin, Jichao Tong, Shiwen Li "Query Optimization of Distributed Database Based on Parallel Genetic Algorithm and Max-Min Ant System" International Symposium on Computational Intelligence and Design, pp. 581-585, 2015.

[5] Vikash Mishra and Vikram SinghGenerating Optimal Query Plans for Distributed Query Processing using Teacher-Learner Based Optimization" International Multi-Conference on Information Processing, pp. 281-290, 2015.

[6] Fazal Mithani, Sahista Machchhar, Fernaz Jasdanwala "A Novel approach for SQL query optimization" 2016 IEEE.

[7] Wazeb Gharibi, Ayman Mousa "Query optimization based on time scheduling" 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 978-1-4799-3080-7/14/ © 2016 IEEE. pp. 1370-1375.

[8] Vivek Shrivastava, Brajesh Patel "An approach to Optimize query using Rank aware scoring function" International conference on computational Intelligence and communication Networks, pp 503-507, 2013.

[9] Nicholas L. Farnan, Adam J. Lee, Panos K. Chrysanthis, and Ting Yu "PAQO: Preference-Aware Query Optimization for Decentralized Database Systems" ICDE Conference, pp. 424-435, 2014

[10] http://searchsqlserver.techtarget.com/definition/database, accessed on 24 October 2017

[11] http://www.geeksforgeeks.org/query-optimization/, accessed on 25 October 2017

[12] https://www.slideshare.net/dixitdavey/query-optimization-10386222, accessed on 26 October 2017