A TIME EFFICIENT CLUSTERING ALGORITHM FOR QUERY OPTIMIZATION IN DISTRIBUTED DATABASE

Juhi Srivastava[1], Prof. Gayatri Pandi (Jain) [2],
Student (Master of Engineering) [1], H.O.D. PG Departments[2],
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 plan based on information available at compilation time. This paper focuses on producing an efficient query execution plan on distributed database. Various Optimization strategies have been reviewed in this paper which are based on genetic algorithm and a new model for query optimization is proposed by replacing FCM (Fuzzy c-mean) clustering algorithm with GFCM (Geostatistical Fuzzy c-mean) clustering algorithm. Simulation result show that proposed algorithm can find optimal query execution plan in a relatively short period of time and thus improve the query efficiency.

Keywords **- Distributed Database, Genetic algorithm, GFCM (Geostatistical Fuzzy c-mean) clustering algorithm, Query, Query Optimization.**

I. 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 is 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 users 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.

1.1 Genetic Algorithms

Genetic Algorithms are the search algorithms based on the mechanics of Natural Selection and Natural Genetics. Potential solutions to the problem are encoded as simple chromosome-like data structure and some recombination operators are applied to them. This activity helps in preserving the important information of the chromosome.

These algorithms are computationally simple yet powerful in their search for improvement. Genetic algorithms are finding more widespread application in business, scientific and engineering circles.

The search in the Genetic algorithm starts with a population of chromosomes (query execution plan). The fitness value of each chromosome in the population, using the fitness function is evaluated. Designing the Fitness function is one of the most crucial aspects of Genetic algorithm as it reflects the "Fitness" or "figure of merits" of a chromosome the fitter individual is then selected for crossover and mutation to arrive at the population for the next generation. GA explores the entire solution space to arrive at an optimal set of chromosomes. Objective function for genetic algorithm used in this paper is defined as follows:

$$f(x) = x(1) * \sin\left(\sqrt{x(1)}\right) - x(2) * \sin\left(\sqrt{x(2)}\right) \qquad (1)$$

1.2 GFCM Clustering algorithm

The fuzzy *C*-means algorithm (FCM) has been utilized in a wide variety of applications. Its advantages include a straightforward implementation, fairly robust behaviour, unsupervised, and the ability to model uncertainty within the data. A major disadvantage of its use is long computation time while clustering the dataset, sensitive to initial guess, sensitive to noise and one expects

to low membership degree for outliers. Therefore, Geostatistical Fuzzy c-mean clustering is used. The advantages of the new method are the following: (1) it yields regions more homogeneous than those of other methods, (2) it removes noisy spots, and (3) it is less sensitive to noise than other techniques. It is derived by extending the FCM objective function.

Here Clustering is a two-pass process at each iteration. The first step is the same as that in standard FCM to calculate the membership function in the spectral domain. In the second step, the membership information of each data point is mapped to the spatial domain, and the spatial function is computed from that. The FCM iteration proceeds with the new membership that is incorporated with the spatial function. The iteration is stopped when the maximum difference between two cluster centers at two successive iterations is less than a threshold. The main aim is to minimize the objective function, where kriging variance is incorporated. It is a derived function.

$$JGF(U,v) = \sum_{i=1}^{N}\sum_{j=1}^{c}(uij)m[d(xi,vj)]2 + \sum_{i=1}^{N}(uij)m\sum_{j=1}^{c}(\text{ej})2 \ -\sum_{i=1}^{n}\lambda i(\sum_{i=1}^{c}uij-1) \qquad (2)$$

where,

The fuzzy clustering of objects is described by a fuzzy matrix μ, with **N rows and c columns** in which **'N'** is the number of data objects and '**c' i**s the number of clusters.

- μ_{ij} [0, 1], i=1, 2…., n, j=1, 2…., c
- $\Sigma\mu_{ij}$ =1, i=1, 2…., n
- j=1 $0<\Sigma\ \mu_{ij}<n$, j=1, 2…., c
- $\mathbf{(e_j)^2}$ is the Geostatistical (kriging) variance of estimating v_j using x_i, i = 1…N-1.

Geostatistical method allow to find a best estimate and its variance. It offers a way of describing the spatial continuity of natural phenomena and provides adaptation of classical regression to take advantage of this continuity. Deals with spatially auto correlated data.

II. LITERATURE SURVEY

ShaoHua Liu, Xing Xu, [1] 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 is 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.

S. Venkata Lakshmi, Dr. Valli Kumari Vatsavayi, [2] propose two different phases. 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.

Manik Sharma, Gurvinder Singh, Rajinder Singh, [3] 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 generates 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.

Wenjiao Ban, Jiming Lin, Jichao Tong, Shiwen Li, [4] 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.

Vikash Mishra and Vikram Singh, [5] the author uses parameter less optimization technique. This algorithm work for multi-objective unconstrained and constrained benchmark. This algorithm considers 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.

III. PROPOSED STRATEGY

The proposed approach models the query optimization problem as a single-objective genetic algorithm where the objective is to minimize the value of objective function.

That is query plan with minimum cost need to be identified.The basic process of genetic algorithm with GFCM clustering algorithm is as follows:

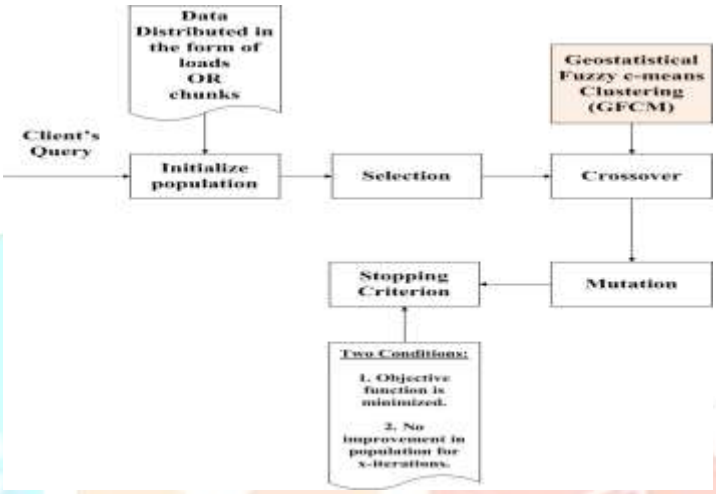

Figure 1: Architecture model of proposed strategy

3.1 Initialize Population

Here data is fragmented and replicated to a number of sites in forms of load and chunks. Then these number of sites are taken to randomly generates query plan as population.

3.2 Selection

From the generated chromosomes (query plan), a query plan with minimum query processing cost is selected as a parent chromosome. Here two types of Selection technique are used:

(i) RWS (Roulette Wheel Selection), in which a fixed point is selected and then the point that comes in front of selected point is selected as parent while rotating the wheel.

(ii) SUS (Stochastic Universal Selection), in which multiple fixed point is selected and from that selected point a new point is selected as a parent.

3.3 Crossover

After Selecting the parent query plan, the database of that query plan are clustered into 10 cluster by using Geostatistical Fuzzy c-mean (GFCM) clustering algorithm which gives more homogeneous environment than Fuzzy c-mean (FCM) clustering algorithm. After data is clustered Crossover operator is applied on the data. Through the crossover operation, a new generation of individual can be obtained, which are combined with the characteristics of their parents. It reflects the idea of information exchange. Here Whole Arithmetic Recombination Crossover operation is used, in which weighted average of the two parents are taken to obtain a new individual.

3.4 Mutation

We randomly select an individual in the group and change the value of one of the string structure data with a certain probability for the selected individual. It provides an opportunity for the generation of new individual. It is used to maintain and introduce diversity in the genetic algorithm. Here Swap Mutation Operator is used in which two position on the chromosome are selected at random and interchange their values.

3.5 Stopping Criterion

There are two simple condition of program stopping:

(i) minimizing the value of objective function, and

(ii) No improvement in population for x-iterations.

IV. SIMULATION RESULT

In order to compare the proposed strategy and improved genetic algorithm, the simulation test was carried out. Experiments were carried out on the PC machine with windows 10 Operating System and Matlab R2014b. Here 7 different databases which are further fragmented and replicated to 28 different sites. In our experiment of distributed database query, upto 50 different query execution plans are generated. The experimental results were carried out in two different criterions:

(i) Number of iteration versus Running time, and

(ii) Range (upper and lower bound) versus objective function. The experimental results are shown in the Figure 2 and 3:

The first experimental result shown in Figure 2, time required for searching the efficient query plan while using GFCM clustering algorithm with genetic algorithm is less than FCM clustering algorithm with genetic algorithm.

Figure 2: Performance analysis in terms of number of iterations versus running time

The second experimental result shown in Figure 3, objective function value for proposed strategy is less than the improved genetic algorithm. Thus distributed query optimization is efficient in proposed strategy.

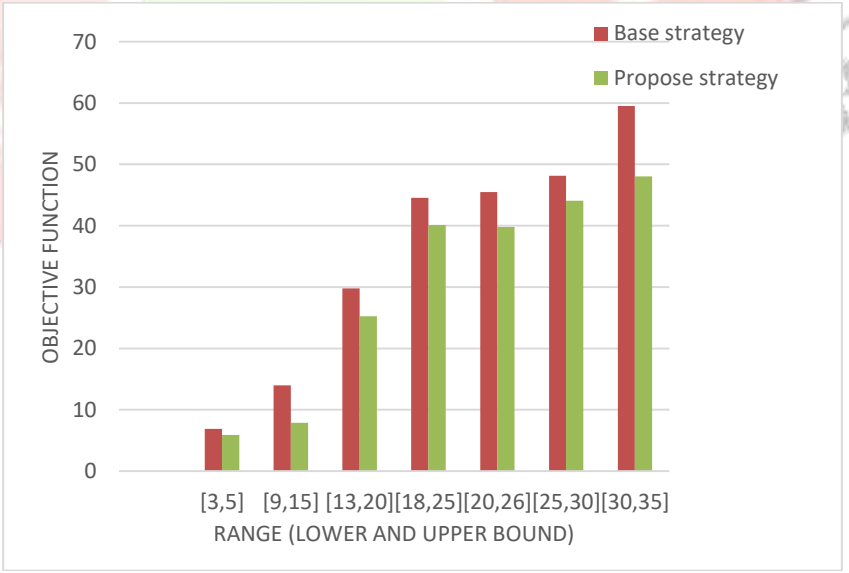

Figure 3: Performance analysis in terms of objective function versus range

Experimental results show that the performance of our proposed strategy is better than improved genetic algorithm and our proposed strategy is more efficient with the increase in number of iteration and range.

V. CONCLUSION

This research paper analyses the shortcoming of improved genetic algorithm in Fuzzy c-mean clustering and proposes a new optimization strategy which replace Fuzzy c-mean clustering algorithm with Geostatistical Fuzzy c-mean. Proposed strategy reduces

the time required to cluster data and thus reduces the overall time required to execute the query. The experimental result show that the proposed strategy is superior to the improved genetic algorithm at the performance of the distributed database query.

VI. FUTURE SCOPE

The Proposed strategy works only on single objective function which can be further improve to multi objective function. That is in future we try to work on multiple objective function while using the same optimization strategy.

VII. ACKNOWLEDGEMENT

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.

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 Optimizer 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 Singh "Generating Optimal Query Plans for Distributed Query Processing using Teacher learner Based Optimization" International MultiConference 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. Chrvsanthis 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