International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© **Global Society of Scientific Research and Researchers**

http://ijcjournal.org/

# Database Optimization Using Genetic Algorithms for Distributed Databases

G. C. A. L. Aponso[a*], T. M. T. I. Tennakon[b], A. M. C. B. Arampath[c], S. Kandeepan[d] , H. P. K. K. S. Amaratunga[e]

[a,b,c,d,e]*Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology, Colombo 03, Sri Lanka*

[a]*Email: anoshaponso@gmail.com*, [b]*Email: tthisurash@gmail.com*

[c]*Email: chinthakaarampath@gmail.com*, [d]*Email: theepan4@gmail.com*

[e]*Email: kasun.mana@gmail.com*

## Abstract

Databases can store a vast amount of information and particular sets of data are accessed via queries which are written in specific interface language such as structured query language (SQL). Database optimization is a process of maximizing the speed and efficiency with which kind of data is retrieved or simply it's a mechanism that reduces database systems response time. Query optimization is one of the major functionality in database management systems (DBMS). The purpose of the query optimization is to determine the most efficient and effective way to execute a particular query by considering several query plans such as graphical plans, textual plans and etc. Execution of any particular datasets depends on the capability of the query optimization mechanism to acquire competent query processing approaches. Distributed database system is a collection several interrelated databases which are spread physically across different environments that communicate through a computer network. Inability to obtain an effective query strategy with an efficient accuracy and minimum response time or cost to execute the given query is one of the major key issues of the query optimization in distributed database systems. Further inefficient database compression methods, inefficient query processing, missing indexes, inexact statistics, and deadlocks are furthermore defects. In this paper, it describes the methodologies such as genetic algorithm strategy for distributed database systems so as to execute the query plan. Genetic algorithms are extensively using to solve constrained and unconstrained optimization problems. The genetic algorithms are using three main types of rules such as selection rules, crossover rules, and mutation rules.

***Keywords:*** Query optimization; Database Management Systems (DBMS); Database optimization; Genetic algorithms; Distributed database systems.

------------------------------------------------------------------------

* Corresponding author.

## 1. Introduction

Query optimization is a most imperative mechanism which comes under the database optimizations that should consider. So this had been considered many different standpoints which provide a number of different solutions in each case. The primary aim of the query optimization is to find an optimal query execution plan that can return results back to the particular user efficient and effectively. When generating optimal query execution plan some characteristics of distributed database systems require the impact of few factors such as network model, data allocation and etc [4].

There are so many researches have been carried out to find a solution for the above, but those approaches didn’t show a much of success rate. Some algorithms have a lot of computational errors when the number of relations increases [5]. The approach is taken by combining iterative improvement algorithm and simulated annealing algorithm also failed because it does not provide an optimal solution but it reduces the search cost [6]. The combination of Parallel Genetic Algorithm (GA) and Max-Min Ant System (MMAS) [1], Evolutionary Algorithm [2], Teacher Learner Based Optimization algorithm [3], Entropy-based restricted stochastic query optimizer (ERSQO) [7] are some of the examples for above-mentioned approaches.

This paper is divided into few sections. The abstract will give the overall idea and the view of the paper. A general introduction to the paper is presented under section one. Section two described the background of the core concepts of the research such as related works and etc. Future works are presented in section three and the fourth section provides the conclusion.

## 2. Background

As discussed earlier Database optimization is a most imperative thing that should consider. So this had been considered many different standpoints which provide a number of different solutions in each case. Even though many researchers have finished the massive effort, however, they were not acceptable as a complete solution for database optimization. W. Ban and his colleagues gave an extensive survey about database optimization for distributed databases. They proposed a query optimization hybrid algorithm that combines Genetic Algorithm and Max-Min Ant System (MMAS). Generally, Genetic Algorithm directly makes the best group of individual and replaces the worst solution. In MMAS all are limited to a maximum and a minimum value which above or below the area will be adjusted automatically. And they have mentioned about distributed database according to the size of query relations, in addition, the size of the intermediate results of two join relations also can be used as cost estimation. In this proposed algorithm allocate several parallel processors to improve the execution speed [1]. Although genetic algorithm convergence faster than ant colony algorithm, it is easier to fall into an early-maturing state. Relatively as an enhanced ant colony algorithm, Max-Min ant system has stronger global search ability, but the lack of initial pheromone lead to slower convergence speed in the early stage of generating an optimal solution. Considering that two algorithms can run in parallel and distributed database provide the parallel environment, the parallelization technique can be applied to improve the hybrid algorithm further. For convenience, this calls this new algorithm PGA-MMAS which is the abbreviation of the parallel genetic algorithm and max-min ant system. It is entirely feasible to combine them together. Besides, considering that two algorithms can run in parallel and distributed database provide the parallel environment, the parallelization technique can be applied to improve the hybrid algorithm further. S. Mansha and F. Kamiran proposed a solution Evolutionary Algorithm to generate communication plan. The queries that were input are merged into application graphs. The fitness of each individual is noted for all generations using rank-based selection method. According to them, this shall help to overcome computational delay [2]. Z. Haider et.al has proposed novel feature extraction algorithm which is used Genetic Algorithm and to optimize the output node they proposed to use Artificial Neural Network (ANN). This does not change training process and modify the results. According to them the proposed algorithm reduces the dimensionality by 41.23% and requires less train time. This function depends on the input because weights obtained are constant In their GA is used for to obtain the relevant features which maximize the objective function [8]. Distributed databases are based on partition or replication. V. Mishra and V. Singh built a Teacher Learner Based Optimization algorithm which generates optimal top query plans[3]. It is based on a Genetic algorithm. According to the results, it is effective and has the potential of solving similar design objectives. According to L.Y Ho et.al on Distributed Database System Query Optimization Algorithm Research a new algorithm is designed through the research on query optimization technology, based on a number of optimization algorithms commonly used in distributed query, a new algorithm is designed which can significantly reduce the amount of intermediate result data, effectively reduce the network communication cost, to improve the optimal efficiency. This is a research on producing an algorithm to improve the efficiency of the database by using query optimization, but the research is not focusing on the use of indexing [9]. Distributed Database Query Optimization is achieved thru many complex sub-operations on the Relations, Network Sites, Local Processing Facilities and the Database System itself. R. Gurvinder and Varinder have proposed a stochastic model simulating a Distributed Database environment, and shown benefits of using innovative Genetic Algorithms (GA) for optimizing the sequence of sub-query operations allocation over the Network Site [10]. Many of these sub problems are NP-Hard itself, which makes Distributed Database Query Optimization a very complex and hard process. One of these NP-Hard components is an optimal allocation of various sub-queries to different sites of data distribution. Most of the prevalent solutions take help of Exhaustive Enumeration Techniques, along with the use of innovative heuristics. The efficiency of distributed database system is significantly dependent on the extent of optimality of this execution plan. The process of generating a good query execution strategy involves three phases. First is to find a search space which is a set of alternative execution plans for the query. Second is to build a cost model which can compare costs of different execution plans. Finally, in the third step, they explore a search strategy to find the best possible execution plan using cost model. Before putting any queries to a Distributed Database, one needs to design it according to the needs of an organization [6]. Queries require efficient processing, which mandates devising of optimal query-processing strategies that generate efficient query processing plans for a given distributed query. The number of possible query processing plan grows rapidly with increase in the number of sites used, and relations accessed, by the query. There is a need to generate efficient query processing plans from among all possible query plans. V. Vickram and Ajay have proposed an approach which attempts to generate such query processing plans using a genetic algorithm. The approach generates query plans based on the closeness of data required to answer the user query. The query plans having the required data residing in fewer sites, are considered more efficient and are thus preferred, over query plans having data spread across a large number of sites.When referring to the earlier approaches taken by several experts it’s really hard to overcome the existing issues. Using genetic algorithm is a most suitable approach that can take to solve the above-mentioned issues. In this type of algorithms, then the time taken when obtaining a suitable solution is unique and independent of the search space, because of that particular reason implementing a genetic approach is more suitable and effective when optimizing the queries in distributed databases [13], [14]. [15], [11] and [12]. Genetic algorithms borrow its essential features from natural genetics and these algorithms are stochastic techniques that produce a good quality solution with minimum time complexity [7]. Further, these types of genetic algorithms have the ability to operate on populations of solutions rather than a single solution and normally employ some heuristics like selection, crossover, and mutation to develop better and efficient solutions [13].

## 3. Future works

This research is needed to be improving more in the unique of database and query optimization, such as conducting more deep analysis and evaluation of design entropy based stochastic query optimizer by considering the impact of a variety of selection approaches of genetic algorithms. Furthermore results of each approach can be compared with each other results in very effective and efficient way and can conduct further research on the impact of data allocation to optimize the usage and the process of the system.

## 4. Conclusion

This research paper has presented a generic algorithm for database quarry optimization. The proposed Distributed Query Optimization technique gives a methodology to generate an efficient distributed query processing plan which improves the reply time of user queries. These methodologies achieved the distributed query processing plan generation as a single-objective genetic algorithm problem. At the same time, take advantage of the parallelism of the algorithm itself and distributed database cluster environment to accelerate the algorithm convergence further. Finally, this showed the results are efficiency in the generic algorithm for database quarry optimization.

## References

[1] W. Ban, J. Lin, J. Tong, and S. Li, “Query Optimization of Distributed Database Based on Parallel Genetic Algorithm and Max-Min Ant System,” 2015 8th Int. Symp. Comput. Intell. Des., no. 1, pp. 581–585, 2015.

[2] S. Mansha and F. Kamiran, “Multi-query Optimization in Federated Databases Using Evolutionary Algorithm,” 2015 IEEE 14th Int. Conf. Mach. Learn. Appl., no. 1, pp. 723–726, 2015.

[3] V. Mishra and V. Singh, “Generating Optimal Query Plans for Distributed Query Processing using Teacher-Learner Based Optimization,” Procedia Comput. Sci., vol. 54, pp. 281–290, 2015.

[4] A. Hameurlain and F. Morvan, “Evolution of Query Optimization Methods,” vol. 33, no. 0, pp. 211–242, 2009.

[5] D. Kossmann and K. Stocker, “Iterative Dynamic Programming : A New Class of Query Optimization Algorithms 1 Introduction,” pp. 1–38.

[6] A. K. Giri, “Distributed Query Processing Plan Generation using Iterative Improvement and Simulated Annealing,” pp. 757–762, 2012.

[7] M. Sharma, “Parametric Analysis of Different GA based Distributed DSS Query Optimizer Models,” pp. 148–154, 2016.

[8] Z. Haider, C. Yin, W. Zhang, L. Zhang, M. Yousaf, and N. Ali, “Enhanced Feature Selection Method Based on ANN and GA for Coal Boiler Plants Using Real Time Plant Data,” pp. 7115–7119, 2016.

[9] L.-Y. Ho, M.-J. Hsieh, J.-J. Wu, and P. Liu, “Data Partition Optimization for Column-Family NoSQL Databases,” 2015 IEEE Int. Conf. Smart City/SocialCom/SustainCom, pp. 668–675, 2015.

[10] R. Singh and V. Gurvinder, “Optimizing Access Strategies for a Distributed Database Design using Genetic Fragmentation,” vol. 11, no. 6, pp. 180–183, 2011.

[11] T. V. V. Kumar, V. Singh, and A. K. Verma, “Distributed Query Processing Plans Generation using Genetic Algorithm,” vol. 3, no. 1, 2011.

[12] E. Sevinc and  a. Cosar, “An Evolutionary Genetic Algorithm for Optimization of Distributed Database Queries,” Comput. J., vol. 54, no. 5, pp. 717–725, 2010.

[13] S. Ender, C. Ahmat, “an evolutionary genetic algorithm for optimization of distributed database queries”, The computer journal, 2011.

[14] P. Tiwari, S. V. Chande, “Optimization of Distributed Database Queries Using Hybrids of Ant Colony Optimization Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering, June, 2013.

[15] S. Ender, C. Ahmat, “an evolutionary genetic algorithm for optimization of distributed database queries”, Oxford University Press on behalf of The British Computer Society, 2010