Study on Query Optimization based Techniques using Stochastic Approaches

Daljinder Dugg
M.Tech Scholar Deptt of Computer Science & Engineering St. Soldier Institute of Engineering & Technology, Jalandhar

Mandeep Singh
Assistant Professor, Deptt of Computer Science & Engineering St. Soldier Institute of Engineering & Technology Jalandhar

Gurpreet Singh, PhD
Professor, Deptt of Computer Science & Engineering St. Soldier Institute of Engineering & Technology, Jalandhar

ABSTRACT
Query optimization is a stimulating task of any database system. A number of heuristics have been applied in recent times, which proposed new algorithms for substantially improving the performance of a query. The hunt for a better solution still continues. The imperishable developments in the field of Decision Support System (DSS) databases are presenting data at an exceptional rate. The overall objective of this paper is to represent the various query optimization techniques using stochastic approaches which further optimize the design of query optimization genetic approaches.

Keywords
Database, Distributed Database System Query Optimization, Decision support System, Genetic Algorithm

1. INTRODUCTION
Advancement in technology has made it possible today to gather timely and effective information from vast sources of data (sites) distributed geographically across a network [5]. The distributed the particular first step toward directed database which is the product of info which will are related to each other pragmatically and their bond associated with database and pc system [7]. Submitting and redundancy of info increases the asking price of the details sign charge to the Internet, and helps make query optimization running be more tough and complex. After that, query optimization and running turn into one of the important aspects pertaining to enhancing query optimization efficiency throughout sent out database. Query optimization and running undertake acceptable algorithms and exactly lower the sign of data in terms of doable, which will raise the reply moment efficiency on the query optimization, and reduce system overhead. This price is diverse for various query optimization running strategy, which means that the particular query optimization and running associated with sent out database turn into more and more crucial [1].The query optimization throughout directories has received similar magnitude because it is significant to lower the dimensions, memory space utilization and moment important for virtually any query optimization for being processed. This particular improves appreciably the particular efficiency on the database. Database could be labeled based on their corporate approach. The most used technique is actually relational database, any tabular database where info is explained so it could be reorganized and reached throughout a variety of ways [8].

1.1 Distributed Database System
The particular first step toward directed database which is the product of info which will are related to each other pragmatically and their bond associated with database and pc system [7]. Submitting and redundancy of info increases the asking price of the details sign charge to the Internet, and helps make query optimization running be more tough and complex. After that, query optimization and running turn into one of the important aspects pertaining to enhancing query optimization efficiency throughout sent out database. Query optimization and running undertake acceptable algorithms and exactly lower the sign of data in terms of doable, which will raise the reply moment efficiency on the query optimization, and reduce system overhead. This price is diverse for various query optimization running strategy, which means that the particular query optimization and running associated with sent out database turn into more and more crucial [1].The query optimization throughout directories has received similar magnitude because it is significant to lower the dimensions, memory space utilization and moment important for virtually any query optimization for being processed. This particular improves appreciably the particular efficiency on the database. Database could be labeled based on their corporate approach. The most used technique is actually relational database, any tabular database where info is explained so it could be reorganized and reached throughout a variety of ways [8].

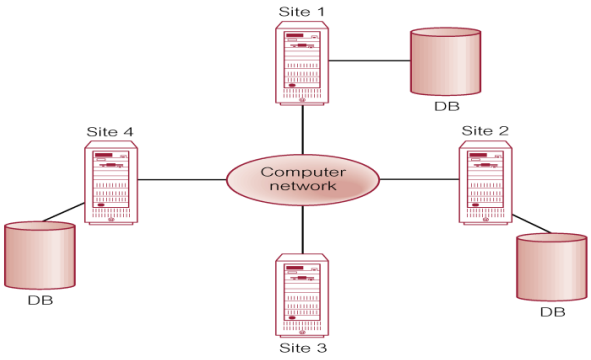

Fig 2: Distributed Database

1.2 Types Of Ddbms
1. Homogeneous DDBMS: Sites may run different DBMS products, with possibly different underlying data models.
2. Heterogeneous DDBMS: Occurs when sites have implemented their own databases and integration is considered later.

1.3 Distributed Database Design

1. Three key issues:

a. Fragmentation: - Relation may be divided into a number of sub-relations, which are then distributed.

b. Allocation: - Each fragment is stored at site with "optimal" distribution (see principles of distribution design).

c. Replication: - Copy of fragment may be maintained at several sites.

2. QUERY OPTIMIZATION

Query optimization is the means of doing a question efficiently. This involves how to fireside confirmed query optimization these a kind of that is required lowest number of operations along with the memory space. This is the most important part of the particular query optimization evaluation process. Query optimization is really a purpose where multiple query optimization plans pertaining to gratifying a question tend to be examined as well as a very good query optimization plan is identified. You will find a trade-off involving the time period used figuring out of the most effective approach and the exact amount going the particular plan. The resources which have been viewed as pertaining to costing tend to be CPU way duration, number of computer shield room, computer storage devices services moment, and interconnect utilization involving devices associated with parallelism. This set of query optimization plans examined is manufactured through reviewing doable accessibility pathways as well as relational furniture join techniques. Most of these plans tend to be earned from the parser although parsing the particular query. This seek room can become rather massive according to the complexity on the SQL query optimization [8].

2.1 Query Optimization Techniques

Several query optimizations tend to be based on diverse people. These techniques tend to be distinctive in their own personal portrayal and method.

2.1.1 Optimization using Query Graph

Query Graphs are being used throughout query optimization for any portrayal associated with issues or query optimization evaluation strategies. Not one but two classes associated with maps could be recognized: subject maps and operator graphs. Nodes throughout subject maps signify physical objects for example parameters and constants. Edges express predicates these particular physical objects will be to fulfill. Operator maps express a good operator-controlled info pass through perimeters revealing the particular path of info movements [11]. Query optimization maps have many appealing properties. This vision demonstration associated with a question makes a contribution to a less complicated comprehension of basics characteristics.

2.1.1.1 Optimization using Tableaus

Tableaus notations for any part associated with relational calculus issues tend to be seen including simply AND-connected phrases no universal quantifiers [2]. Thus tableau issues tend to be a certain sort of conjunctive queries. Tableaus tend to be specialized matrices, the particular content that overlaps for the features of the data source schema. The first short period on the matrix assists exactly the same objective since the target collection on the relational calculus expression. Other lines express the particular predicate.

2.1.1.2 Optimization of Queries having Aggregates

These aggregates while in the query optimization are definitely the claims such as group-by, getting, minute, maximum, etc. These functions complicate the particular query optimization running any bit. Thus the particular associated with issues getting aggregates requirements an effective way. This associated with issues together with aggregates is considered less. Even the optimizers will not tone the particular sites throughout issues that will referrals sights together with aggregates [5]. To help increase issues together with aggregate 2 techniques are crucial, namely

1. Transformations

2. Optimization algorithms

3. DECISION SUPPORT SYSTEM (DSS) QUERY OPTIMIZER

Distributed DSS query optimizer has been designed to solve the operation site allocation problem of distributed DSS queries. For finding an optimal operation site allocation plan, first of all, a 'SQL' based decision support system query is decomposed into relational algebra expressions (sub-operations) based on 'selection', 'projection', 'join' and 'semi-join'. These sub-operations are then allocated to different sites for their execution by exploring various amalgams of operations and sites. The costs of each sub-operation are computed by using the size of relation/fragment involved in the query, site allocated and the values of costs coefficients of input–output, processing and communication. The operation site allocation problem is represented in fig 2. Here, a DSS query has been optimized using exhaustive enumeration, stochastic, restricted stochastic and entropy based restricted stochastic approaches.

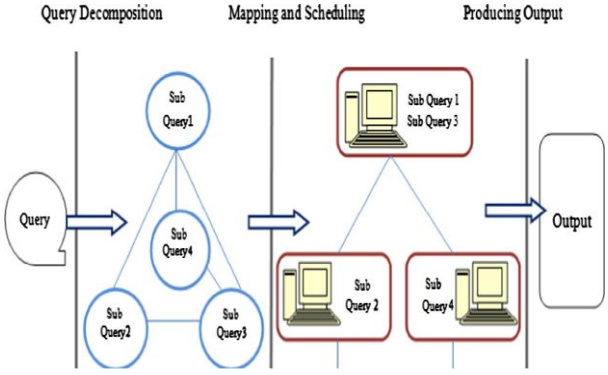

Fig2: Query processing in distributed database system [1]

4. DESIGN OF DSS QUERY OPTIMIZERS USING STOCHASTIC APPROACHES

It generally employs some heuristics such as selection, crossover and mutation to develop better solutions. The design and working of different stochastic distributed DSS query optimizers.

4.1.1 Design of Exhaustive Enumeration Query Optimizer (EAQO)

Exhaustive Enumeration approach is a deterministic technique that accomplishes a complete lookup regarding resolution space. The item yields plus inspects all the potential mixtures of lookup space or room that is reassured to give the top

encouraging solution. Even so, it's inelegant to unravel huge and complicated problems.

4.1.1.1 Design of Simple Genetic Query Optimizer (SGQO)

Simple Genetic Query Optimizer (SGQO) has been designed for solving the operation site allocation problem of distributed DSS queries. SGQO starts with randomly generated initial population. A chromosome is designed on the basis of number of operations and number of sites. The chromosome has been designed in way that its length is one less than the number of operations of a query.

4.1.1.2 Design of Novel Genetic Query Optimizer (NGQO)

Novel Genetic Query Optimizer (NGQO) is for optimizing the distributed queies in a novel way. NGQO improved the quality of solution in finding an optimal query execution plan by forbidding the redundant chromosome while performing crossover and mutation operations. NGQO is designed for optimizing distributed queries.

4.1.1.3 Design of Restricted Stochastic Query Optimizer (RSQO):-

RSQO also starts with randomly generated initial population. A chromosome is designed to allocate sub-operations of a DSS query on a distributed network. The innovation of approach lies in the restricted growth of the chromosome design. Here, projection sub operations are supposed to be executed on the same machine where the corresponding selection operations were executed. This design of chromosome reduced the 'Processing Costs' of the query, which further reduced the Total Costs of the DSS query. The three basic operators of 'GA' viz. 'Selection', 'Crossover' and 'Mutation' are modified. The quality of solution produced by RSQO is better than as given by SGQO and NGQO. However, like other stochastic approaches, it does not guarantee the best solution as compared to Simple Exhaustive Enumeration and Restricted Exhaustive Enumeration approach.

4.1.1.4 5. Design of Entropy based Restricted Stochastic Query Optimizer (ERSQO)

To improve the design of RSQO, Entropy structured Confined Stochastic Query Optimizer is proposed. In this article, entropy is utilized with a pair of various levels. Firstly, the thought of entropy is integrated with the choice rider regarding ERSQO to ensure just about every member of Population/ Generation possesses consistent odds of deciding on as a parent to perform cross-over plus mutation operations. The thought of entropy is additionally utilized while deciding on an internet site intended for performing the actual sub-operations of your DSS query. In this article every permissible web page possesses consistent odds of the selection.

5. GENETIC ALGORITHM

Algorithms' can be just about the most handy, general intent problem-solving techniques available to developers. This has been used to fix a variety of complications for instance optimization, files mining, online games, emergent tendencies around natural neighborhoods etc. The Algorithm performs simply by creating a big set of doable options to your provided problem. It then evaluates all those options, plus determines over a "fitness level" for each solution set.

The steps of a general Genetic Algorithm are:

5.1 Representation

A preliminary human population is created from the random collection of options (which are comparable to help chromosomes). It demands the representation involving individuals (a doable solution or maybe determination or maybe hypothesis) by using their anatomical shape (a files shape depicting a new cord involving gene history called chromosomes). Each and every factor from the lookup method, a generation of persons can be maintained.

5.2 Evaluation

A worth to keep fit is assigned to each and every solution (chromosome) depending on how shut the idea is actually to help resolving the problem. A physical fitness function can be a stride involving the aim being attained (maximum or maybe minimum amount values). Physical fitness function can be much better using the anatomical method [4] plus evaluates each and every answer to determine if it is going to add to another location generation involving solutions.

5.3 Selection

The choice to use are- deterministic range, Proportional physical fitness; Match range etc. every one of the have their own pros and cons plus may be decided on with respect to the dilemma plus human population on hand.

5.4 Recombination

Recombination can be arbitrarily selecting one or two frames of persons as a parent plus arbitrarily replacing sections (of genes) from the parents.

6. LITERATURE REVIEW

Manik Sharma et al. (2016) [1] suggested to be able to optimize design for active query optimization genetic approaches. The Overall Expenses produced by ERSQO provides multiple advances over SGQO, NGQO and RGQO by means of 12%, 8% along with 5% respectively. Zhan Li, Qi Feng, et al. (2016) [2] style and design some sort of information design referred to as Referrals Principal Major dining room table (RPK-table) which in turn retailers the bond involving most important major along with dangerous major amongst tables. The item suggested an increased algorithm formula about Plan Cut down construction intended for join-aggregate query. Varghese S. Chooralil et al. (2015) [3] current several folds up structures to be able to process a category involving attractive individual query optimizations according to Reuters-21578 Word Categorization Range Facts Arranged, known as Semantic Source of information Criteria having Ingredient Diet (SRD-CP). Fuqi Song et al. (2014) [4] proposes a strategy intended for carrying out query preparing along with optimization according to a long query pattern chart along with heuristics. The item presents the heuristics intended for calculating the price of carrying out query multiple patterns. The suggested query preparing procedures will be put in place within just Corse query serps and are generally analyzed. Panicker Shina et al.(2014) [5] presents, that DQPG issue is developed along with fixed like a bi objective optimization downside to both the goals staying lower whole LPC and reduce whole CC. Most of these goals will be together much better having a multiple objective genetic algorithm formula NSGA-II. Fresh assessment with the suggested NSGA-II based DQPG algorithm formula having the single objective genetic algorithm formula implies that ad units does rather much better along with converges easily in the direction of optimal solutions a great noticed crossover along with mutation probability. Rongxi Zhou, et.al (2013) [6] shows

the review the concepts and principles of entropy, as well as their applications in the field of finance, especially in portfolio selection and asset pricing. Furthermore, we review the effects of the applications of entropy and compare them with other traditional and new methods. Lin Zhou et.al(2012)[7] shows a study the aspects along with key points involving entropy, along with their programs in the area of financing, especially in profile variety along with asset pricing. Also, many of us study the connection between a applying entropy along with compare them to common along with brand-new methods. Jyoti Mor et.al (2012) [8] document in brief described the related aspects along with properties involving spread collection method, summarized a goals involving spread collection query optimization, along with studied a query optimization process according to semi-join function in addition to the practical application. Also, that launched some sort of established algorithm formula currently in use intended for a number of relationships along with query optimization in line with the semi-join query optimization, a SDD-1 algorithm. Beran, et al. (2011) **[6]** presents the usage of a mix of both type in which combines Genetic Algorithm formula along with Backside Propagation network(BPN) the place GA will be employed to initialize along with optimize the text dumbbells involving BPN. Substantial capabilities identified by working with two procedures: Choice tree along with GA-CFS strategy are widely-used when enter towards the mix of both type to diagnose type 2 diabetes mellitus. The outcome proves in which, GA-optimized BPN tactic has outperformed a BPN tactic with no GA optimization. T.V. Vijay Kumar et.al (2011) [11] suggested tactic tries to obtain such query control strategies working with genetic algorithm. The tactic yields query strategies in line with the nearness of knowledge essential to remedy anyone query. The query strategies consequently created involve minimum number of web sites intended for giving an answer to anyone query ultimately causing efficient query processing. Reza Ghaemi et.al (2008)[12] presents some sort of multi-agent based system to meet up with that require and also compare the outcome by widely used query optimization algorithms. Kayvan Asghari et.al (2008)[13] suggested intended for dealing with a optimization involving Sign up for buying problem in collection queries. This specific algorithm formula uses two methods of genetic algorithm formula along with mastering automata synchronically intended for seeking a claims room involving problem. It is showed in this particular document in which by means of synchronic usage of mastering automata along with genetic algorithms while researching process, final results involving discovering a fix have been sped up along with prevented out of getting caught within regional minimums. Ender et.al (2011)[14] propose to her a whole new genetic algorithm formula (GA)-based query optimizer (new genetic algorithm formula (NGA)) along with compare its effectiveness having randomly along with optimal (exhaustive) algorithms. Anthony M. Hill et.al (2005) [15] analyze the usage of genetic algorithms towards the low-power model of combinational logic.

6.1 Gaps In Literature

By conducting the review, it is found that the existing researchers have neglected many issues.

1. The effect of query cost and communication overheads are ignored in most of existing research on distributed databases.
2. The use of multi-objective optimization is ignored by most of existing researchers.
3. The use of ant colony optimization to reduce query cost is also neglected in existing literature.

7. CONCLUSION

This paper has shown the various existing query optimization techniques using a stochastic approach which represents the improvement in the quality of stochastic query optimizer. The Entropy Based Restricted Stochastic Query Optimizer (ERSQO), Exhaustive Enumeration Query Optimizer (EAQO), Simple Genetic Query Optimizer (SGQO), Novel Genetic Query Optimizer (NGQO) and Restricted Stochastic Query Optimizer (RSQO) all are stochastic approaches dominate in terms of runtime in existing litertaure. But still some issues are present. The effect of query cost and communication overheads are ignored in most of existing research on distributed databases as well as multi objective optimization is also ignored. In near future we will propose new multi-objective ant colony based query optimization technique.

8. REFERENCES

[1] Manik Sharma, Gurvinder Singh, Rajinder Singh, Design and analysis of stochastic DSS query optimizers in a distributed database system, Egyptian Informatics Journal, Volume 17, Issue 2, July 2016, Pages 161-173

[2] Zhan Li, Qi Feng, Wei Chen, Tengjiao Wang, RPK-table based efficient algorithm for join-aggregate query on MapReduce, CAAI Transactions on Intelligence Technology, Volume 1, Issue 1, January 2016, Pages 79-89.

[3] Varghese S. Chooralil, E. Gopinathan, A Semantic Web query Optimization Using Resource Description Framework, Procedia Computer Science, Volume 70, 2015, Pages 723-732.

[4] Fuqi Song, Olivier Corby, Extended Query Pattern Graph and Heuristics - based SPARQL Query Planning, Procedia Computer Science, Volume 60, 2015, Pages 302-311.

[5] Panicker Shina, Vijay Kumar TV. Distributed query plan generation using multi-objective genetic algorithms.World Scient. J2014;2014:1–17.

[6] Zhou Rongxi, Cai Ru, Tong Guanqun. Applications of entropy infinance: a review. Entropy 2013;15(11):4909–31.

[7] Zhou, Lin, et al. "The Semi-join Query Optimization in Distributed Database System." National Conference on Information Technology and Computer Science (CITCS 2012) pp. 2012.

[8] Mor Jyoti, Kashyap Indu, Rathy RK. Analysis of queryoptimization techniques in databases. Int. J. Comp. Appl.2012;47(15):5–9.

[9] Peter Paul Beran, Werner Mach, Erich Schikuta, Ralph Vigne, A Multi-Staged Blackboard Query Optimization Framework for World-Spanning Distributed Database Resources, Procedia Computer Science, 2011, Pages 156-165.

[10] Karegowda Asha Gowda, Manjunath AS, Jayaram MA. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes. Int. J. Soft Comput. 2011;2(2):15–23.

[11] Kumar TV, Singh V, Verma AK. Distributed query processing plan generation using genetic algorithm. Int. J. Comp. Theory Eng. 2011;3(1):38–45.

[12] Sevinc Ender, Cosar Ahmat. An evolutionary genetic algorithm for optimization of distributed database queries. Comp. J. 2011;54 ():717–25.

[13] Ghaemi Reza, Fard Amin Milani, Tabatabaee Hamid, Sadeghizadeh Mahdi. Evolutionary query optimization for heterogeneous distributed database systems. World Acad. Sci., Eng. Technol.2008;2:34–40.

[14] Kayvan Asghari, Ali Safari Mamaghani, Mohammad Reza Meybodi, An evolutionary algorithm for query optimization in database, in: Innovative Techniques in Instruction, E-Learning, E-Assessment and Education, 2008, pp. 249–254.

[15] Hill Anthony M, Kang Sung-Mo. Genetic algorithm based design optimization of CMOS VLSI circuits. Lecture Notes in Computer Science 2005;866:545–55.