ISSN 2278-3091

Volume 8, No.5, September - October 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse14852019.pdf

https://doi.org/10.30534/ijatcse/2019/14852019

Database Query Optimization using Genetic Algorithms: A Systematic Literature Review

Swati V. Chande

International School of Informatics and Management, Jaipur, India, swatichande@rediffmail.com

ABSTRACT

As data is escalating exponentially efficient storage and fast retrieval of data are increasingly becoming areas of significance for researchers. Query optimization is probably the most researched problem in the database domain. Among the solutions to this problem, Genetic Algorithms have proved to be a viable alternative. This systematic literature review explores work done on the application of Genetic Algorithm to optimization of database queries to understand how its use for database query optimization has taken shape and to identify the open questions. It could be inferred that Join-Order, Distributed databases and Hybrids with Genetic Algorithm are sustaining researchers' interest.

Key words: Data Management Systems, Query Optimization, Genetic Algorithms.

1. INTRODUCTION

Over the decades making the results of a query available to a user at the earliest has been the focus of research. Given a query, there are a lot of alternative plans that a database management system can use to process the query and retrieve the right response for the user. All these plans, called query evaluation plans, are the same in terms of the final results they produce, but differ a lot in terms of their cost, or duration they require to execute, and the intermediate storage that they may require. However, with storage becoming less expensive, it is hardly a constraint and hence, optimization of a query primarily involves searching for the plan that may need the minimum, or about minimum amount of time to execute the query [1].

The query optimization problem is of enormous research interest within the database field. In the paper [2], the author has given a lucid description of the query optimization problem. The problem has been studied by several researchers in diverse perspectives and using different approaches, resulting in quite a few solutions in each case. Some of these studies include [3, 4, 5, 6].

Several algorithms have been tried and studied to optimize a database query [7]. Some of these algorithms have proved to be a viable alternative to existing algorithms and have performed even better for large join queries. The solutions however have not yet seen much inclusion in the standard database management systems. Genetic Algorithms are one such solution.

Since the 1991 paper by Bennett et. al. [8] on the use of Genetic algorithms for query optimization, several researchers have studied this problem. Genetic Algorithms have been among the most extensively researched optimization algorithms for database query optimization. For the research being carried out in the domain for over quarter of a century now, it is time to know how effective it has been, and that is what this study aims to do.

This work aims to explore the literature on the application of Genetic Algorithms to optimization of database queries to understand how the use of GA for database query optimization has taken shape, and to identify the issues and open questions for further studies. Additionally, this study particularly attempts to find out the dimensions studied by researchers in the domain and to explore whether the researchers in database query optimization field have focused consistently on application of Genetic Algorithms to the problem.

The Systematic Literature Review approach, guided by identified research questions using the population, intervention, comparison, outcome, and context (PICOC) criteria forms the methodology for carrying out this study.

More than 3500 research studies published since 1991 were surveyed, subsequently the most significant approaches were selected, and in them, the Genetic Algorithms based solutions to the database query optimization problem was comprehensively reviewed. We thus traced the course that research in the GA based database query optimization field has taken since 1991, and also the continuity of research, implementation, challenges, and open questions in the field.

The findings of the survey are based on a significant degree of coverage concerning the techniques for database query optimization using genetic algorithms. This work also reveals the scope and directions that the studies have taken in the course of research in over two decades, and the future direction the studies may take.

Genetic Algorithm (GA) and Systematic Literature Review (SLR) are thus the key elements that constitute this study of the evolution of the GA based solutions to the database query optimization problem.

Genetic Algorithms

A genetic algorithm is a bio-inspired search algorithm based on the concept of natural selection and survival of the fittest. It evolves to search for the fittest solutions from a large population of alternative solutions to a problem. Genetic algorithms have since long been used for optimization of large join queries, searching for a plan with near-least cost for execution of a query. As queries increase in size and complexity, current algorithms seem inadequate to cater to the increasing demands of searching an optimal path for query execution.

SLR

A systematic literature review (or systematic review facilitates identification, evaluation and interpretation of existing studies related to a specific research question, or subject, or domain of concern [9]. SLRs facilitate creation of a summary of current evidence that can contribute to evidence-based furthering of research in the area of interest. This is made possible due to reduction of large amount of research information into lucid units, and aggregation of critical information for analysis by way of a systematically laid out plan. For this study, the area of interest is query optimization in databases using genetic algorithms.

Area of interest

Efficient techniques of optimizing complex queries are a key requirement for effective performance of database management systems. With continuous increase in size and complexity of data, contemporary query optimization techniques seem insufficient to aid some of the evolving database applications. Researchers therefore are continuously on the job for searching techniques to better the performance of queries and reduce response time. Several algorithms have been considered for this purpose. Genetic Algorithms have been studied extensively as initial studies have revealed their viability for efficient query optimization for large join queries.

Over the last twenty five years or so, researchers have studied application of different aspects and forms of genetic algorithms to solve the database query optimization problem. A systematic analysis of the research done so far on this could provide a comprehensive view of the directions and dimensions covered by the studies, extent of applicability and implementation, and probable future moves.

In this study, a systematic review is carried out to summarize the existing evidence on the subject of application of Genetic Algorithms to optimization of database queries. This review also identifies gaps in current research in the domain, and suggests quarters for further investigation. It also provides an outline and background so as to suitably position new research activities.

2. METHODOLOGY

Planning a well thought out methodology is necessary for conducting a systematic literature review. The following subsections discuss the methodology adopted for this study.

2.1 Study Design

This systematic literature review provides a broad outline of the research carried out in the area of optimization of database queries using genetic algorithm, establishes whether sufficient research substantiation is available on the topic, and provides quantitative proof of the same, as per the guidelines/ specifications described in [9, 10]. The literature review aims to summarize the work on DBQO using GA and identify probable directions for further studies.

The review is conducted following commonly accepted practical guidelines [8, 9] to plan the study. The study protocol available in the studies by [11] and [12] and used in [13] was used in this study.

The systematic literature review was done by following the steps given below.

Research questions—identify and state the research questions to form the basis of the study

Search strategy— plan working scheme and identify the resources searched to compile data, and the search string

Article selection— state conditions to select or reject the studies

Chronology of studies— plan chronological criteria for inclusion of studies

Quality assessment— set criteria for assessment of quality of the selected studies

Data extraction— view how the studies are addressing the research questions, and what data to consider for analysis.

The sections below describe how the process was carried out as per the design of the study.

A. Research Questions

As per [9] and [10], stating of research questions (RQ) is the most critical step of a systematic review. Identification and classification of, i) the studies in the area concerned, ii) the characteristics, complications, challenges, and resolutions being considered in the existing studies, and iii) the existing or emerging research prospects, has hence been attempted. Considering the same, general and specific research questions have been framed. The general research questions have been refined into more specific questions for providing an classification and thematic analysis, and also for exhaustive determining likely research dimensions for further analysis. The research questions have been categorized as, general question (GQ) and specific question (SQ). Table 1 includes the research questions considered.

Table 1: Research Questions

Identifier	Question
GQ 1	How has the research on use of GA for database query optimization taken shape?
GQ 2	Has it succeeded in retaining continuous focus of the researchers in database query optimization?
GQ 3	What are the challenges and open questions related to Genetic Algorithms based database query optimization?
SQ 1	What modifications are done in the techniques involving Genetic Algorithms for improving the performance of the query optimizers?
SQ 2	Have these studies seen practical implementations outside the laboratories in any databases?

GQ 1 focuses on the natural twists and turns the research in the domain must have taken to be in sync with the changes in the database field. GQ 2 refers to the relevance of the applicability of Genetic Algorithms to the database query optimization problem during the period of study. GQ 3 is to look into the possible future prospects, issues, and course of action in research in the field. With the general research questions, certain consequent specific research questions (SQs) to identify specific issues and techniques used for GA based database query optimization have also been defined. These questions have been proposed to identify questions and solutions adjoining genetic query optimization. SQ1 answers the questions regarding modifications in the solutions comprising Genetic Algorithms to improve performance of query optimizers. SQ 2 tries to find out how well have the solutions proposed by researchers been accepted by the software systems developers.

B. *Search Strategy*

The subsequent phase involves finding a comprehensive compilation of research publications associated with the research questions. The step involves selection of search keywords and definition of search scope [10]. The keywords were identified to get precise search results. Authors in [9] have suggested splitting the research question into discrete aspects as research units, where their synonyms, abbreviations, and alternative spellings are all incorporated and integrated using Boolean operators. In [10] authors have proposed the PICOC (population, intervention, comparison, outcome, and context) criteria that form a guideline for suitable framing of research questions. The PICOC criteria for this review are defined as given below and presented in Table 2.

Population

The population involves keywords, related terms, variants, synonyms for the issue under consideration. To ensure that the search string is finalized after considering all likely words, the 'word similarity' tool was used for the basic technical terms, namely, 'Database', 'Query', and 'Optimization'. The tool identifies synonyms, related terms and variants of the submitted term. The synonyms of the terms were thus obtained.

It was however observed that being technical terms, these terms are used as they are, and not their variants. 'Genetic Algorithm' being the specific algorithm it would be used as it is in the papers. In this study, since all aspects of studies carried out in the database query optimization field are to be explored, not many restrictions and constraints were required. It was therefore decided that these terms would be searched in the entire document and not just in the title.

Therefore, the following search string is defined for the selection.

> "Database" and "Query Optimization" and "Genetic Algorithm"

Intervention

To filter studies effectively as per the requirements, studies which described optimization by genetic algorithms or hybrids involving genetic algorithms only were considered.

Comparison

The comparison phase was not included in the study.

Outcomes

The outcomes related to effectiveness of genetic query optimization in databases.

Context

Review of all studies involving Genetic Algorithms for optimization of database queries.

Table 2: PICOC criteria

Phase	Description
Population	GA based QO solutions for databases
Intervention	GA or hybrid of GA for QO in database
Comparison	N/A
Outcomes	Effectiveness of GA
Context	Reviews of studies involving GA for QO in database

C. *Article Selection*

Once all the related articles were found, the next step was to eliminate the not so relevant papers so as to retain only the most representative ones. The studies not addressing GA based database query optimization specifically, were hence removed. For applying the inclusion/exclusion criteria, population and intervention criteria were used as follows:

Exclusion criterion 1: Paper includes the terms but does not involve specific study on genetic query optimization in databases.

Exclusion criterion 2: Studies involving use of Genetic Algorithms for Data Mining and Data Warehousing were not included.

Inclusion criteria 1: Paper includes Genetic Algorithm at least as one of the components of the solution to the database query optimization problem.

The steps of the filtering process are:

(1) impurity removal,
(2) filter by title,
(3) filter by keyword,
(4) filter by abstract,
(5) elimination of duplicates, and
(6) filter by full text.

As a first step, the impurities in the search results were eliminated. Certain impurities, such as, names of conferences, journals, workshops, seminars, bibliographies, author databases etc. had the search keyword and, were included in the search results, these were removed.

Next, the titles of the articles were considered and the ones that did not have database query optimization as the subject were excluded.

This was followed by filtering on the basis of keywords. Articles that did not have subject relevant keywords were dropped. Then, the abstracts of the articles were analyzed and the ones that did not relate to the topic under consideration were excluded.

Next, the studies that remained after the application of the above filters were congregated and replicas were dropped as some studies occurred in over one electronic database and some were repeated in the same database.

After the application of filtering criteria some studies not particularly relevant to this survey remained among the selected ones. These were removed after analyzing their full text.

D. *Chronology of Studies*

Studies carried out since 1991 were considered for inclusion.

E. *Quality Assessment*

It is necessary to ensure that the selected studies fulfill some quality criteria. The quality measure is used to validate that the research publication is indeed a significant contribution [10]. The articles were evaluated on the objective of research, its relevance and context, review of related work, methodology used, results obtained, conclusion being in line with the stated objectives, as well as pointing out of future studies. The criteria for evaluation of quality of selected studies are given in Table 3.

Table 3: Quality Assessment Criteria

Identifier	Issue
C1	Is the objective of the research stated unambiguously?
C2	Is the literature review adequate and suitably analyze the literature?
C3	Is literature reviewed fittingly related to the primary subject under consideration?
C4	Is the research methodology appropriately described?
C5	Is the study generating results?
C6	Is the paper presenting a conclusion in line with the research objectives?
C7	Is the paper able to suggest future works based on its contributions, improvements in techniques and methodologies, or further studies?

F. *Data Extraction*

An evaluation form was created to collect particulars of the research papers and their sections to obtain answers to general as well as specific research questions as depicted in Table 4. It maps the sections of the study to the research questions they are likely to answer.

Table 4: Data Extraction Format

Section	Description	Research questions
Open content		
Title	Title of the study	GQ1, SQ1
Abstract	Summary of paper, briefly stating its objective, methodology, and results	GQ1, GQ3, GQ4, SQ1, SQ2
Keywords	Keywords demonstrating the focus of the study	GQ1, SQ1
Article content		
Introduction and Review of Literature	A brief of the entire study could be obtained particularly, the problem under consideration and the theories and prior work associated with it	All questions
Method	The Methodology of carrying out the study	All questions
Results	Evaluation of the work done in the study in accordance with the proposed methodology	All questions
Conclusion	Outcomes as per the stated objectives and	All questions

Section		Description	Research questions
		scope of future work	

3. RESULTS

This section presents the results obtained after the evaluation of full studies related to the research topic. Each proposed research question is answered in the ensuing subsections. Consequently, apart from finding answers to the research questions, the study has proposed contributions made in the genetic algorithm based database query optimization field from the study of related works.

The results obtained vis-a-vis the study protocol set above are given here.

3.1 Research Questions

Answers to the research questions identified are presented at the end of section 3 after taking into consideration all other steps in the protocol.

3.2 Search Strategy

The PICOC criterion was applied to identify the search string. The electronic databases searched to collect the research studies for analysis, 10 in number, were selected to include as many related studies as possible. A list of the same is given in Table 5.

These databases include relevant journals and conferences in the Computer Science field. As mentioned above, duplicate studies retrieved from same or different databases were eliminated by manual filtering during the article selection phase.

Table 5: Electronic Databases/ Search Engines used

Acronym	Full Name
IEEE	Institute of Electrical and Electronics Engineers
ACM	Association for Computing Machinery
MS Academic	Microsoft Academic
DBLP	Digital Bibliography & Library Project
Google Scholar	Google Scholar
ScienceDirect	ScienceDirect
Wiley Online Library	Wiley Online Library
Semantic Scholar	Semantic Scholar
SpringerLink	SpringerLink
CiteSeer^x	CiteSeer^x

3.3 Article Selection

After applying the inclusion / exclusion and filtering criteria given in 2.1.3 relevant studies were selected. The process of selection of the studies is described below.

The initial search prior to application of the exclusion criteria had 3679 articles, of these, 246 (6.63%) articles were identified as impurities. After withdrawal of these articles, the filter by title criterion was applied on the remaining studies. Continuing the process, 2724/3433 (79.35%) publications were filtered by way of a title review, and 467/709 (65.87%) articles were filtered through keyword review. Further 75/ 242 (30.99%) studies got filtered after abstract scrutiny. The remaining publications were grouped, and 109/167 (34.73%) of them were found to be duplicates and therefore dropped. Subsequently after the application of exclusion criterion 2 to the full text, only 60/107 (44.95%) studies remained.

Some of the selected 60 studies, were by the same author(s) and quite similar. The most recent/ representative research paper from among these was chosen as the methodology and tools more or less remained the same in them. Thus, 21/60 (35%) publications were eliminated. Consequently, 39 publications remained selected for this study as the reference line.

Further, to analyze the progress and the recent trends, these articles were divided into two sets S1 and S2 representing the time period. S1 includes 16 studies conducted between 1991 and 2007, and S2 includes 23 papers published in the last decade (2008-2018).

The studies in S1 are more reflective of the foundations of the database query optimization using genetic algorithms field, they were therefore considered to answer the RQs, GQ 1, SQ 1 and SQ 2.

An overview of the 16 studies in S1 is given in Table 6 (a) with the identifier representing the set and research study number, author reference, publication year, publisher, and publication type, which are sorted in ascending order by publication year.

Table 6 a: Articles in S1

Identifier	Author(s), Year of Publication	Publisher	Type
S1RS01	Bennett et al, 1991 [8]	University of Wisconsin	Technical Report
S1RS02	Jorng-Tzong Horng et al, 1994 [14]	IEEE	Conference
S1RS03	Sang Koo Seo et al, 1996 [15]	Elsevier	Journal
S1RS04	M. Utesch, 1997 [16]	PostgreSQL	PostgreSQL Programmer's Guide
S1RS05	Steinbrunn	ACM	Journal

Identifier	Author(s), Year of Publication	Publisher	Type
	M. et al, 1997 [7]		
S1RS06	Nafjan K.A. et al, 1997 [17]	Springer	Conference
S1RS07	Rho, S. et al, 1997 [18]	Springer	Journal
S1RS08	Sushil J. Louis et al, 1998 [19]	Association for the Advancement of Artificial Intelligence (AAAI)	Conference
S1RS09	M. Gregory, 1998 [20]	IEEE	Conference
S1RS10	Ikeji A.C. et al, 1998 [21]	Springer	Conference
S1RS11	Ahmad, I. et al, 2002 [22]	Springer	Journal
S1RS12	Varga, Viorica et al, 2004 [23]	Springer	Conference
S1RS13	Ali Safari Mamaghani et al, 2007 [24]	World Scientific and Engineering Academy and Society (WSEAS)	Conference
S1RS14	Murat Ali Bayir et al, 2007 [25]	IEEE	Journal
S1RS15	H Dong et al, 2007 [26]	ACM	Conference
S1RS16	Zhou Z, 2010 [27]	World Academic Press (WAP)	Journal

An overview of the 23 studies of S2 is given in Table 6 (b) with the identifier representing the set and research study number, author reference, publication year, publisher, and publication type, which are sorted in ascending order by publication year.

Table 6 b: Articles in S2

Identifier	Study, Year	Publisher	Type
S2RS01	Hongxing Li et al, 2008 [28]	IEEE	Conference
S2RS02	Stoyan Vellev, 2008 [29]	Bulgarian Academy of Sciences	Journal
S2RS03	Kayvan Asghari et al, 2008 [30]	Springer	Conference
S2RS04	Suk-Kyu Song, 2009 [31]	Wiley	Journal
S2RS05	Ender Sevinç et al, 2009 [32]	IEEE	Conference
S2RS06	T. V. Vijay Kumar et al, 2010 [33]	IEEE	Conference
S2RS07	Najmeh Danesh et al, 2010 [34]	Academic Journals	Journal
S2RS08	M. Sinha et al, 2010 [35]	Academic Journals	Journal
S2RS09	Gorla, Narasimhaiah et al, 2010 [36]	JCS&IT	Conference
S2RS10	Swati V. Chande et al, 2011 [37]	IEEE	Conference
S2RS11	H. Kadkhodaei et al, 2012 [38]	IEEE	Conference
S2RS12	Tansel Dökeroğlu et al, 2012 [39]	VLDB Conference	Conference
S2RS13	Tiwari, P. et al, 2013 [40]	Advance Academic Publisher	Journal
S2RS14	Singh V et al, 2014 [41]	ACM	Conference
S2RS15	Gajjam, N S et al, 2014 [42]	Foundation of Computer Science	Journal
S2RS16	M. Sharma et al, 2015 [43]	IEEE	Conference
S2RS17	Sambit Kumar Mishra et al, 2015 [44]	Science and Education Publishing	Journal
S2RS18	Wenjiao B et al, 2015 [45]	IEEE	Conference
S2RS19	S. Liu et al, 2016 [46]	IEEE	Conference
S2RS20	Mishra V, 2016 [47]	Springer	Conference
S2RS21	Ming Y, 2017 [48]	IEEE	Conference
S2RS22	Shiwu Y, 2017 [49]	DEStech Publications	Conference
S2RS23	Venkata L, 2017 [50]	-	Journal

3.4 Chronology of Studies

Studies since 1991 were considered for inclusion. A view, as depicted in Figure 1, at these studies reveals that there has been a consistent interest of the researchers in the field.

Particularly from 2006 onwards, there has been no gap in publications in the domain.

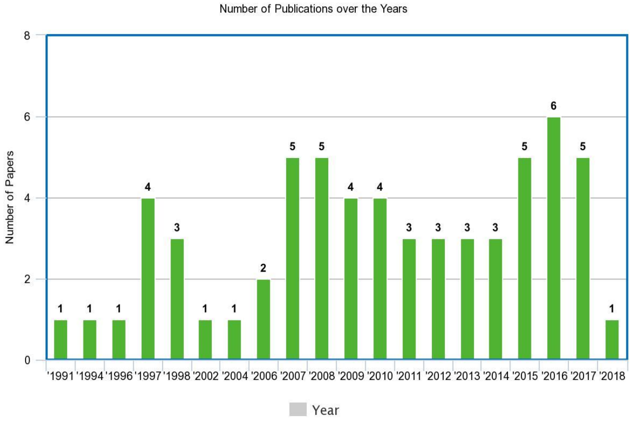

Figure 1: Chronology Graph: Number of Publications per year.

3.5 Quality Assessment

Figures 2a and 2b, illustrate the score on the quality measures of the articles in S1 and S2 respectively on the basis of quality assessment criteria proposed in Table 3.

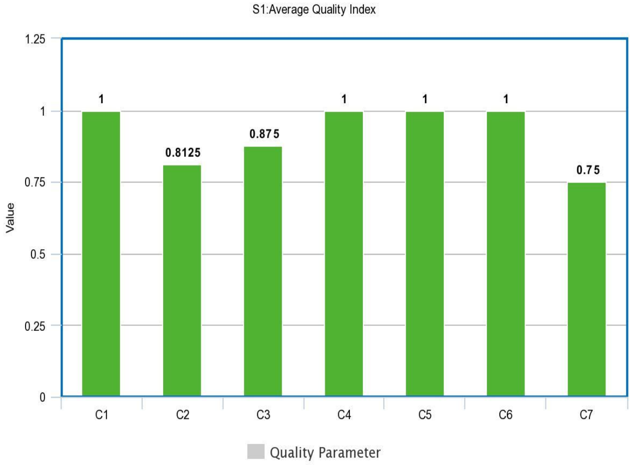

Figure 2a: Quality assessment of articles in S1

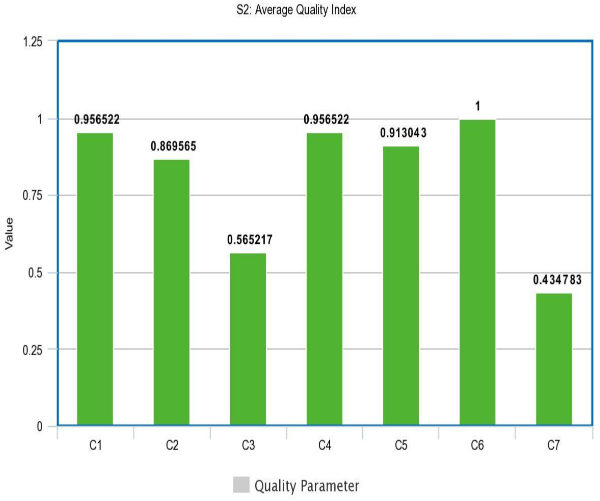

Figure 2b: Quality assessment of articles in S2

The average quality criteria score of the articles is shown on the vertical axis and the criteria on the horizontal axis. The publications scored well on the criteria for evaluation, recording a minimum of 4 points out of the total of 7 points considering each fulfilled criteria to be one point. For example, many articles do not include possible future studies as they are conclusive studies.

3.6 Data Extraction

The research papers were reviewed section-wise as per the format given in Table 4. It helped in directing the analysis to precise points that could answer the research questions identified.

3.7 Answers to the Research Questions

Finally, the research questions, both GQs and SQs, have found the answers as given below. First, the answers to GQs,

GQ 1: How has the research on use of GA for database query optimization taken shape?

Since the paper by Bennett et al in 1991 on application of genetic algorithm to database query optimization which focused on join order optimization for queries in centralized databases, the area has attracted considerable attention. The shift has naturally been towards distributed databases to be in sync with the changing database landscape. Researchers have also attempted use of hybrid algorithms with GA as one of the constituents. Of the 60 studies reviewed after exclusion criterion 39 were applied for full text filter. These 39 studies were divided in to two sets, S1 and S2 chronologically, 16 pre-2008 in S1 and 23 studies carried out since 2008 in S2. These studies reveal broadening of the spectrum of application of GA to the query optimization problem as depicted in Table 7 and Figure 3.

The studies in S1 continued focus on Join order optimization. Of the 16 studies, 25% were on use of GA for query optimization in distributed databases indicating a clear interest in the evolution of the subject and an effort to apply it to the upcoming changes in the database field. Two of the studies involved application of a hybrid algorithm with GA as a component. Researchers have also suggested new operators giving extension to the query language, modifications in GA operators, and new heuristics in GA based database query optimization. This initial period thus, in a way set the tone for things to come distributed databases, hybrid algorithms, and modifications in Genetic Algorithms,

In S2, Join Order Optimization for large queries continues to be the primary area of research. Distributed databases take the center stage with about 70% of the studies dedicated to them. The other key domains of interest include Hybrid Algorithms and Modified Genetic Algorithms. GA has been hybridized with several algorithms, and it itself has undergone modifications of different types in these studies. Researchers have studied query proximity and rank aware queries.

Table 7: Domains of research identified

Group	**Number of Studies**
Hybrid	7
Modified GA	7
Query Proximity	6
Total Time Versus Response Time	2
Rank Aware Queries	1

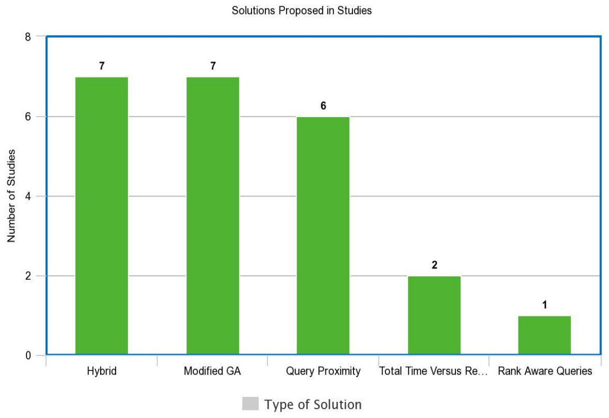

Figure 3: Solutions proposed and number of studies

The research on use of GA for database query optimization thus has been able to establish viability in the initial phase, and in the last decade, it has gained depth by way of researchers delving deeper into application of GA in Database Query Optimization.

GQ 2: Has it succeeded in retaining continuous focus of the researchers in database query optimization?

There has been continuity in research on the use of genetic algorithms for query optimization in databases. As can be seen in the graph in Figure 1 above, researchers have followed GA based solution as a viable solution for database query optimization. Particularly from 2006 onwards, there have been regular publications in the field. Of the 60 studies shortlisted after the full text passed the exclusion criterion 2, 48 have been carried out between 2006 and 2018. Use of GA in DBQO therefore is an ongoing study and one which is sustaining researchers' attention since almost three decades.

GQ 3: What are the challenges and open questions related to Genetic Algorithms based database query optimization?

The GA for database query optimization problem is very well a thriving research domain. On the basis of the studies reviewed, it can be deduced that researchers are finding GA to be a practicable solution for query optimization in Genetic Algorithms. The field has several dimensions which further studies can take. These include,

Use of modified mutation and crossover operators (S2RS22)

Use of adaptive mutation and crossover operators (S2RS22)

Exploring use of hybrid algorithms with GA as a component while also incorporating certain domain-specific heuristics and randomized local search techniques (S2RS22)

Implementing GA for fuzzy queries (S2RS30)

Use of Natural Language Processing, including horizontal and vertical fragmentation of the tables (S2RS43)

But this paper ignores the cost of CPU and I/O which may influence the multi-join query of distributed database to a certain extent, we will address this issue as a future work. (S2RS56)

Application of Intelligent Approaches (S2RS59)

And now, we have the answers to SQs.

SQ 1: What modifications are done in the techniques involving Genetic Algorithms for improving the performance of the query optimizers?

Researchers have applied several techniques involving Genetic Algorithms to improve the performance of database query optimizers as summarized in Table 7. Table 8 elaborates the techniques used.

Table 8: Modifications proposed by researchers for effective GA based database query optimizers

Group	**Specialty**	**#**
Hybrid	ACO	3
Hybrid	TLA	1
Hybrid	Immune Theory	1
Hybrid	FCM	1
Hybrid	Learning Automata	1
Modified GA	Tree Based GA	1
Modified GA	Variations in Crossover Operator	2
Modified GA	Vector Evaluation GA	1
Modified GA	Adaptive GA	1
Modified GA	Parallel GA	1
Modified GA	Entropy Based	1
Query Proximity	No. of Relations & Sites containing them	3
Query Proximity	Query Proximity Cost	3
Total Time Versus Response Time	Total Time Versus Response Time	2
Rank Aware Queries	Rank Aware Queries	1

Alluvial diagrams help to visualize nature of variation in a group. They depict significant changes which can be further highlighted using different colors, and ease recognition of key transitions. The RAWGraphs tool introduced in [51] is used here to illustrate by way of alluvial diagrams, in figure 4, the nature of variation that has taken place in the research in GA based database query optimization field.

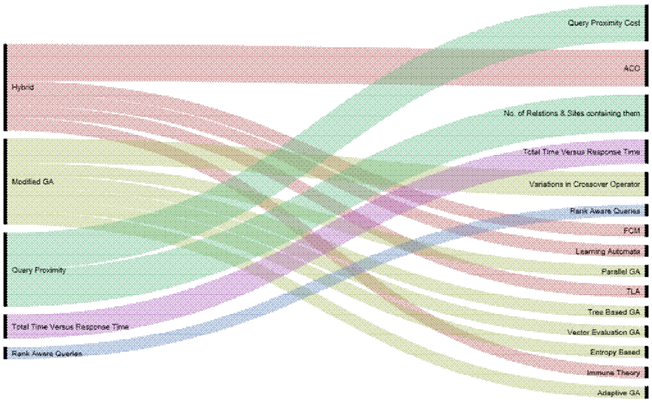

Figure 4: Modifications proposed by researchers for effective GA based database query optimizers

SQ 2: Have these studies seen practical implementations outside the laboratories in any databases?

Of the papers studied only two, S1RS04 and S2RS10 mentioned practical application in a database, of which S1RS04 has been incorporated as a component of the PostgreSQL database.

4. LIMITATIONS

This study attempted to answer the identified research questions so as to chart an outline of the existing literature related to database query optimization using genetic algorithms and its variants. The research was restricted to searching publications in scientific portals related to Computer Science and Information Technology. Only the publications obtained from these websites after application of steps of systematic literature review methodology, were included. The study only considered research papers and did not include publications that were generic or commercial in nature.

5. CONCLUSION

This study was carried out to identify and deliberate on the core concerns apropos database query optimization and trace progress of solutions to the problem. In order to find answers to the RQs of the study, the information obtained from the prior studies on the subject was systematized, quantified, and qualified and it subsequently served as a source for review and analysis.

The review revealed the issues and challenges in application of Genetic Algorithms to optimization of database queries. It brought forward the wide range of solutions proposed by researchers and commonalities in their solutions. It could be inferred that Join-Order, Distributed databases and Hybrids with GA are sustaining researchers' interest. The ever increasing size and complexity of databases and database queries is motivating further research in this domain.

Aspects about concerns and challenges in the adoption of Genetic Algorithms based solutions in DBMS applications could be identified. In addition to answering the research questions, the answers and categorizations found contribute to identification of future scope of research in the field. In future studies, a focus on upcoming database requirements in the context of query optimization can be envisioned. Though these questions have been present since quite some time, they have not found conclusive answers so far. Also there is continuous research going on in the field of optimization algorithms suggesting modifications in existing algorithms and/or their hybrids such as [52], [53] etc. and this would further encourage studies of their application to database query optimization.

REFERENCES

1. Y. E. Ioannidis. **Query optimization**, *ACM Computing Survey*, Vol. 28, No. 1, March 1996, pp. 121-123. https://doi.org/10.1145/234313.234367
2. S. Chaudhuri. **An overview of query optimization in relational systems**, Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (PODS '98), ACM, New York, 1998, pp. 34-43.
3. L. F. Mackert, and G. M. Lohman. **R* optimizer validation and performance evaluation for local queries,** *Proceedings of the 1986 ACM SIGMOD international conference on Management of data*, ACM, New York, 1986, pp. 84-95.
4. L. F. Mackert, and G. M. Lohman. **R* Optimizer Validation and Performance Evaluation for Distributed Queries,** *Proceedings of the 12th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.*, San Francisco, 1986, pp. 149-159.
5. W. Du, R. Krishnamurthy, and M. C. Shan. **Query optimization in heterogeneous DBMS**, *Proceedings of the Conference on Very Large Data Bases,* Vancouver, Canada, August 1992, pp. 277– 291.
6. M. Stillger, G. M. Lohman, V. Markl, and M. Kandil. **LEO - DB2's LEarning Optimizer**, *Proceedings of the 27th International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., San Francisco, 2001, pp. 19-28.
7. M. Steinbrunn, G. Moerkotte, and A. Kemper. **Heuristic and randomized optimization for the join ordering problem**, *The VLDB Journal* Vol. 6, No. 3, August 1997, pp. 191-208.
8. K. Bennett, M. C. Ferris, and Y. Ioannidis. **A genetic algorithm for database query optimization**, *Proceedings of the Fourth International Conference on Genetic Algorithms,* 1991, pp. 400-407.
9. B. Kitchenham, and S. Charters. **Guidelines for performing systematic literature reviews in software engineering**, version 2.3., EBSE Technical Report EBSE-2007-01, Software Engineering Group, School of

Computer Science and Mathematics, Keele University, UK and Department of Computer Science, University of Durham, UK, 2007.

10. M. Petticrew, and R. Helen. ***Systematic Reviews in the Social Sciences: A Practical Guide***, Blackwell Publishing, 2005.
11. J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos. ***Systematic review in software engineering***. Tech. Rep. RT-ES 679/05, System engineering and computer science department COPPE/UFRJ, 2005.
12. D. Qiu, B. Li, S. Ji, and H. Leung. **Regression Testing of Web Service: A Systematic Mapping Study**, *ACM Computing Survey*, Vol. 47, No. 2, Article 21, August 2014, pp. 46.
13. A. Roehrs, C. A. da Costa, R. D. Righi, and K. S. de Oliveira. **Personal Health Records: A Systematic Literature Review**. *Journal of Medical Internet research*, Vol. 19, No. 1, 2017.
14. H. Jorng-Tzong, K. Cheng-Yan, and L Baw-Jhiune. **A genetic algorithm for database query optimization**, *Proceedings of the First IEEE Conference on Evolutionary Computation*, *IEEE World Congress on Computational Intelligence*, Orlando, 1994, pp. 350-355 vol.1.
15. S. K. Seo, and Y. J. Lee. **Applicability of genetic algorithms to optimal evaluation of path predicates in object-oriented queries**, *Information Processing Letters, Elsevier BV*, Vol. 58, Issue 3, 1996, pp. 123-128.
16. M. Utesch. **Genetic query optimization in database systems**, *PostgreSQL Programmer's Guide*, 1997.
17. K. A. Nafjan, and J. M. Kerridge. **Large join order optimization on parallel shared-nothing database machines using genetic algorithms**, *Proceedings of Parallel Processing, Euro-Par 1997, Lecture Notes in Computer Science Springer*, Vol. 1300, Berlin, Heidelberg, 1997.
18. S. Rho, and S. T. March. **Optimizing distributed join queries: A genetic algorithmic approach**, *Annals of Operations Research*, Vol. 71, 1997, pp. 199-228.
19. S. J. Louis, and Y. Zhang. **An Empirical Comparison of Randomized Algorithms for Large Join Query Optimization**, *Proceedings of the Eleventh International FLAIRS Conference*, 1998, pp. 95-100.
20. M. Gregory. **Genetic algorithm optimisation of distributed database queries**, IEEE International Conference on Evolutionary Computation Proceedings, IEEE World Congress on Computational Intelligence, Anchorage, 1998, pp. 271-276.
21. A. C. Ikeji, and F. Fotouhi. **Optimization of constrained queries with a hybrid genetic algorithm**, *Database and Expert Systems Applications DEXA, Lecture Notes in Computer Science, Springer*, Vol. 1460, Berlin, Heidelberg, 1998.
22. I. Ahmad, K. Karlapalem, Y. K. Kwok, and S. K. So. **Evolutionary Algorithms for Allocating Data in Distributed Database Systems**, *Distributed and Parallel Databases, Springer*, Vol. 11, Issue 5, 2002, pp. 5-32.
23. V. Varga, D. Dumitrescu, and C. Groşan. **Solving Stochastic Optimization in Distributed Databases Using Genetic Algorithms**, *Advances in Databases and Information Systems, Lecture Notes in Computer Science, Springer*, Vol. 3255, Berlin, Heidelberg, 2004.
24. A. S. Mamaghani, K. Asghari, F. Mahmoudi, and M. R. Meybodi. **A novel hybrid algorithm for join ordering problem in database queries**, *Proceedings of the 6th WSEAS international conference on Computational intelligence, man-machine systems and cybernetics, World Scientific and Engineering Academy and Society (WSEAS)*, Stevens Point, USA, 2007, 104-109.
25. M. A. Bayir, I. H. Toroslu, and A. Cosar. **Genetic Algorithm for the Multiple-Query Optimization Problem**, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 37, No. 1, pp. 147-153, Jan. 2007.
26. H. Dong, and Y. Liang. **Genetic algorithms for large join query optimization**, *Proceedings of the 9th annual conference on Genetic and evolutionary computation (GECCO '07), ACM*, New York, USA, 2007, pp. 1211-1218.
27. Z. Zhou. **Using Heuristics and Genetic Algorithms for Large-scale Database Query Optimization**, *Journal of Information and Computing Science*, Vol. 2, 2007, pp. 261-280.
28. H. Li, and B. Luo. **A Tree-based genetic algorithm for distributed database**, *IEEE International Conference on Automation and Logistics*, Qingdzao, 2008, pp. 2614-2618.
29. S. Vellev. **Review of Algorithms for the Join Ordering Problems in Database Query Optimization**. *International Book Series Information Technologies and Control, Bulgarian Academy of Science*, 2009, pp. 82-88.
30. K. Asghari, A. S. Mamaghani, and M. R. Meybodi. **An Evolutionary Algorithm for Query Optimization in Database**, *Proceedings of Conference on Innovative Techniques in Instruction Technology, E-learning, E-assessment, and Education. Springer*, Dordrecht, 2008.
31. S. K. Song. **A Genetic Algorithm for Minimizing Query Processing Time in Distributed Database Design: Total Time Versus Response Time**, *The Kips Transactions, Journal of the Korea Information Processing Society*, Volume 3, Part 16D, 2009, pp. 295-306.
32. E. Sevinc and A. Cosar. **An evolutionary genetic algorithm for optimization of distributed database queries**, *Proceedings of 24th International Symposium on Computer and Information Sciences*, Guzelyurt, 2009, pp. 147-152.
33. T. V. V. Kumar, V. Singh and A. K. Verma. **Generating Distributed Query Processing Plans Using Genetic Algorithm**, *Proceedings of the International Conference on Data Storage and Data Engineering*, Bangalore, 2010, pp. 173-177.

34. N. Danesh, H. Shirgahi, H. Motameni. **Optimizing N relations join queries by genetic algorithm**, *Scientific Research and Essays,* Vol. 5, No. 13, 2010, pp. 1576-1582.
35. M. Sinha, and S.V. Chande. **Query Optimization Using Genetic Algorithms**, *Research Journal of Information Technology*, 2010, pp. 139-144.
36. N. Gorla, and S. K. Song. **Subquery allocations in distributed databases using genetic algorithms**, *J Comput Sci Technol*, Vol. 10, 2010, pp. 31-37.
37. S. V. Chande, and M. Sinha. **Genetic optimization for the join ordering problem of database queries**, *Annual IEEE India Conference*, Hyderabad, 2011, pp. 1-5.
38. H. Kadkhodaei, and F. Mahmoudi. **A combination method for join ordering problem in relational databases using genetic algorithm and ant colony**. *Proceedings of the IEEE International Conference on Granular Computing*, 2011, pp. 312-317.
39. T. Dökeroğlu. Parallel Genetic Algorithms for the Optimization of Multi-Way Chain Join Queries of Distributed Databases, Very Large Databases conference, 2012.
40. P. Tiwari, and 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*, Vol. 3, No. 6, June - 2013, pp. 609-614.
41. V. Singh, and V. Mishra. **Distributed Query Plan generation using Aggregation based Multi-Objective Genetic Algorithm**, *Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, ACM*, New York, USA, 2014. Article 26, 8 pages.
42. N. S. Gajjam, and S. S. Apte. **Reducing Execution Time of Distributed SELECT Query in Heterogeneous Distributed Database using Genetic Algorithm**, *International Journal of Computer Applications*, Vol. 100, No. 7, 2014.
43. M. Sharma, G. Singh, R. Singh and J. Singh. **Design and Analysis of Stochastic Query Optimizer for Biobank Databases**, *15th International Conference on Computational Science and Its Applications*, Banff, 2015, pp. 47-51.
44. S. K. Mishra, S. Pattnaik, and D. Patnaik. **Estimation of Fitness Parameter along with Weight of Query Plans in Distributed Database Environment Using Genetic Algorithm Techniques**, *American Journal of Computing Research Repository*, Vol. 3, No. 2, 2015, pp. 14-17.
45. W. Ban, J. Lin, J. Tong, and S. Li. **Query Optimization of Distributed Database Based on Parallel Genetic Algorithm and Max-Min Ant System**, *8th International Symposium on Computational Intelligence and Design IEEE*, 2015.
46. S. Liu, and X. Xu. **Distributed Database Query Based on Improved Genetic Algorithm**, *3rd International Conference on Information Science and Control Engineering (ICISCE)*, Beijing, 2016, pp. 348-351.
47. V. Mishra, and V. Singh. **Vector Evaluated Genetic Algorithm-Based Distributed Query Plan Generation in Distributed Database**, *Proceedings of the International Conference on Recent Cognizance in Wireless Communication & Image Processing, Springer*, New Delhi, 2016.
48. M. Yao. **A distributed database query optimization method based on genetic algorithm and immune theory**, *8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, 2017, pp. 762-765.
49. S. Ye, and Y. Peng. **An Optimization for Distributed Database Multi-join Query Based on Improved Genetic Algorithm**, *Proceedings of 3rd International Conference on Electronic Information Technology and Intellectualization (ICEITI)*, 2017.
50. V. S. Lakshmi, and V. K. Vatsavayi, **Teacher-learner & multi-objective genetic algorithm based query optimization approach for heterogeneous distributed database systems**, *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 8, 2017, pp. 1797-1807.
51. M. Mauri, T. Elli, G. Caviglia, G. Uboldi, and M. Azzi, **RAWGraphs: A Visualisation Platform to Create Open Outputs**, *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, ACM, New York, USA, 2017, pp. 28:1–28:5.
52. M. Zemzami, N. Elhami, M. Itmi, and N. Hmina, **An evolutionary hybrid algorithm for complex optimization problems**, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No. 2, 2019, pp. 126 – 133. https://doi.org/10.30534/ijatcse/2019/05822019
53. M. Zemzami, N. Elhami, M. Itmi, and N. Hmina, **Interoperability Optimization using a modified PSO algorithm**, International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No. 2, 2019, pp. 101-107. https://doi.org/10.30534/ijatcse/2019/01822019