IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

An Empirical Performance Evaluation of Relational Keyword Search Techniques

Madhuri V. Diwan[1], Sangve S.M[2]

M.E Computer (Engineering), Zeal Society's Zeal's College of Engineering and Research, Narhe, Savitribai Phule Pune University, Pune, India[1, 2]

Abstract: Number of approaches has proposed for relation keyword search but for the evaluation of proposed techniques there remains several lack of standardization. There are more irregularity has been in results from different techniques of relational keyword search. Observing more no of results we reach at there are lacks of technology transfer coupled with discrepancies between existing system indication is that there is need for thorough, independent empirical evaluation of search techniques. In this paper, we present the most extensive empirical performance evaluations of relational keyword search techniques to appear to date in the literature. Our results indicates that most of existing techniques not giving acceptable performance for realistic retrieval tasks. Memory consumption precludes more search techniques from scaling beyond small data sets. We also explore the relationship between execution time and factors varied in previous evaluations; analysis of this indicates that these factors relatively little impact on performance.

Keyword: Relational database; data mining; database queries; keyword search; information retrieval; ranking; keyword search.

I. INTRODUCTION

Mining of data is a process of studying data in order to bring about trends or patterns from the data. There are the different number of techniques which is used for mining like text mining, data mining also web mining. In mining process number of algorithms used from that algorithms clustering is one of the most important mining algorithm and which is used for grouping of similar objects together. Clustering technique used for organizing the objects into meaningful manner that makes easy for further process analysis. It is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Thus, for organizing objects into groups in such way that they are similar to each other in some way which is the methodology. The ubiquitous search text box has transformed the way people interact with information. Nearly half of all Internet users use a search engine daily performing in excess of 4 billion searches. The success of keyword search stems from what it does not require namely, a specialized query language or knowledge of the underlying structure of the data. Internet users increasingly demand keyword search interfaces for accessing information, and it is natural to extend this paradigm to relational data. This extension has been an active area of research throughout the past decade. Despite a significant number of research papers being published in this area, no research prototypes have transitioned from proof of concept implementations into deployed systems.

The implicit assumption of keyword searches that is, the search terms are related complicates the search process because typically there are many possible relationships between search terms. It is frequently possible to include another occurrence of a search term by adding tuples to an existing result. This realization leads to tension between the compactness (and consequently performance) and coverage of search results. Composing coherent search results from discrete tuples is the primary reason that searching relational data is significantly more complex than searching unstructured text. Unstructured text allows indexing information at the same granularity as the desired results. This task is impractical for relational data because an index over logical (or materialized) views is often an order of magnitude larger than the original data.

II. RELATED WORK

What Are We Searching For? Analyzing User Objectives When Searching Relational Data. (2012). AUTHOR NAME: J. Coffman and A.C. Weaver

In this paper, we investigate the types of queries that users might submit to a relational keyword search system. We develop a framework for analyzing queries in this context, which is similar to existing taxonomies for web searches. Our analysis of search engine query logs reveals considerable variation among queries for different datasets. We show how to use our framework to create representative query workloads to evaluate relational keyword search techniques and illustrate the differences between our work loads and those used previously in the literature. This work closes an important gap in the existing evaluation methodology and promotes continued improvement to relational keyword search techniques.
In addition, we found that relatively few queries reference more than one database entity unless the additional entity

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

is used to disambiguate the query. When we examined previous evaluations of relational keyword search systems, we found that most evaluations contain more complex search tasks than are present in existing Internet search engine logs. Finally, we demonstrate how to use query templates to construct a representative query workload that is consistent with our analysis of user queries. Such work would allow our analysis techniques to be used for other datasets that we did not consider; it would also serve to validate our existing results by comparing them to the classification of a much larger set of queries.

2.Providing Built-in Keyword Search Capabilities in RDBMS (2011)
AUTHOR NAME: G. Li, J. Feng, X. Zhou, and J. Wang

In this paper we propose a new concept called Compact Steiner Tree (CS Tree), which can be used to approximate the Steiner tree problem for answering top-k keyword queries efficiently. We propose a novel structure-aware index, together with an effective ranking mechanism for fast, progressive and accurate retrieval of top-k highest ranked CS Trees. The proposed techniques can be implemented using a standard relational RDBMS to benefit from its indexing and query-processing capability. We have implemented our techniques in MYSQL, which can provide built-in keyword-search capabilities using SQL. The experimental results show a significant improvement in both search efficiency and result quality comparing to existing state-of-the-art approaches. However, the existing RDBMS technologies have not been designed with supporting keyword search in mind; and the keyword search methods recently proposed by the database community are also largely independent of the underlying RDBMS. While the advantage for keyword-based search to be supported by the underlying RDBMS is quite clear, this integration task remains to be an open challenge. It is further complicated by other constraints such as user friendliness (using keyword search, not SQL queries) and no modification of any source code of an existing RDBMS.
This novel indexing method can be seamlessly incorporated into any existing RDBMS, without the need to modify the source code of RDBMS. It can achieve a good performance by using the capabilities of the underlying RDBMS to support keyword- based search in relational databases. Our proposed approach has been implemented in MYSQL. The experimental results confirm that our approach can achieve high efficiency and result quality, and significantly outperforms state-of-the-art methods.

3. Evaluating the Effectiveness of Keyword Search (2010).
AUTHOR NAME: W. Webber

In this paper, we examine the evolving practices and resources for effectiveness evaluation of keyword searches on relational databases. We compare practices with the longer-standing full-text evaluation methodologies in information retrieval. In the light of this comparison, we make some suggestions for the future development of the art in evaluating keyword search effectiveness.
Keyword search on unstructured text data has long been studied in the information retrieval community, where it goes under the name of free text search. Keyword searches provide only an approximate specification of the information items to be retrieved. Therefore, the correctness of the retrieval cannot be formally verified, as it can with query languages such as SQL. Instead, retrieval effectiveness is measured by user perception and experience. The empirical assessment of keyword-based retrieval systems is therefore imperative.
Effectiveness of a response to a keyword query, and hence of the similarity metric, is not something that can be formally proved. Keyword search offers a straightforward, intuitive, and flexible method of retrieving information.

4.Toward Scalable Keyword Search over Relational Data (2010).
AUTHOR NAME: A. Baid, I. Rae, J. Li, A. Doan, and J. Naughton.

In this paper we argue that as today's users have been "spoiled" by the performance of Internet search engines, KWS systems should return whatever answers they can produce quickly and then provide users with options for exploring any portion of the answer space not covered by these answers. Our basic idea is to produce answers that can be generated quickly as in today's KWS systems, then to show users query forms that characterize the unexplored portion of the answer space.
Combining KWS systems with forms allows us to bypass the performance problems inherent to KWS without compromising query coverage. We provide a proof of concept for this proposed approach, and discuss the challenges encountered in building this hybrid system. Finally, we present experiments over real-world datasets to demonstrate the feasibility of the proposed solution.
The traditional keyword search generate all the answers it can within some time bound, and to augment the search with a form-based approach that "covers" potential answers that the keyword search could not find in the specified time limit. Results from experiments with this approach indicate that it is successful in always returning a covering combination of answers and forms in a bounded and predictable amount of time. We regard this work as a first step toward building this kind of system, and hope that it is a springboard for follow-on work that improves the performance and quality of such systems. In general, exploring the trade-offs between the form-based component and the keyword-based component is fertile ground for future work. For example, a form that returns no answers when executed can convey information to the user about facts that are not present in the database, which is something that seems difficult to capture with a pure keyword-based approach.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

5.Ranking Support for Keyword Search on Structured Data Using Relevance Models (2011)
AUTHOR NAME: V. Bicer, T. Tran, and R. Nedkov

In particular, we adopt relevance-based language models to consider the structure and semantics of keyword search results, and introduce novel strategies for smoothing probabilities in this structured data setting. Using a standardized evaluation framework, we show that our work largely and consistently outperforms all existing systems across datasets and various information needs.
Keyword search on structured data is a popular problem for which various solutions exist. We focus on the aspect of keyword search result ranking, providing a principled approach that employs language models to capture results, queries and the relevance behind them.
A recent study has shown that existing heuristics and normalizations proposed for this problem exhibit good results only in the previous ad-hoc experiments, but fail to deliver consistent performance across different information needs and datasets, and especially, do not deliver stable performance across the precision-recall curve (low precision at higher recall levels).
Through a standardized evaluation, we show that our approach delivers superior results, largely outperforming all existing systems in terms of precision, recall and MAP. Further, we formally show that the ranking function is monotonic. This is of great value in practice, enabling the proposed ranking scheme to be used in combination with state of the art approaches for the efficient computation of results.

III. PROPOSED WORK

Existing evaluations of relational keyword search techniques are ad hoc with little standardization. Webber summarizes existing evaluations with regards to search effectiveness. Our previous work compares relational keyword search techniques with regard to search effectiveness but does not consider runtime performance.
Many existing keyword search techniques have unpredictable performance due to unacceptable response times or fail to produce results even after exhausting memory. We propose a system to overcome the disadvantages which discussed for efficient keyword search. Data mining or information retrieval. One important advantages of keyword search is user does not require a proper knowledge of database queries. User easily inserts a keyword for searching and gets a result related to that keyword.
This working is for explaining about the performance and search effectiveness in the evaluation of large number of search techniques in efficient manner. Many existing search techniques do not provide acceptable performance for realistic retrieval tasks. Intention of proposed system is to avoid all the disadvantages of existing systems. Combine number of algorithms and techniques from data structure and introduce new techniques that can satisfy number of expectations for keyword query search. Proposed System generates the more number of queries for selected keyword. Clustering algorithm is used in proposed system. For our evaluation, we use the DBLP11 data set, which we decomposed into relations according to the schema shown. Y is an instance of a conference in a particular year. PP is a relation that describes each paper pid2 cited by a paper pid1, while PA lists the authors aid of each paper pid. Notice that the two arrows from P to PP denote primary-to foreign- key connections from pid to pid1 and from pid to pid2.

IV. ARCHITECTURAL VIEW

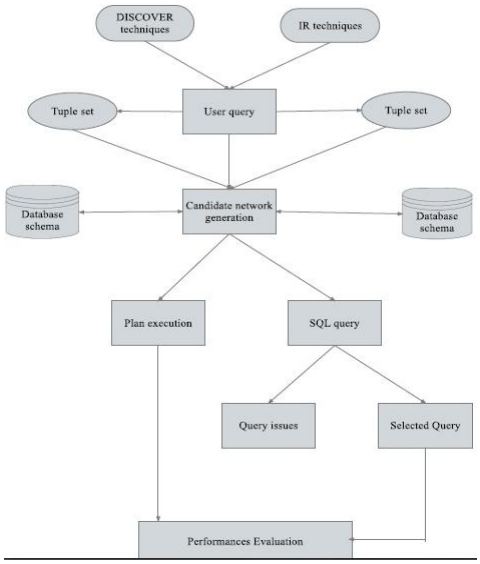


Figure: System Architecture

Our results naturally depend upon our evaluation benchmark. While we would like our experiments to include additional test collections created by other researchers, our benchmark is currently the only publicly available collection of data sets, queries, and relevance judgments one implementation issue that does impact the results for the graph-based search techniques is the graph data structure. All the implementations in our evaluation use the JGraphT library, 15 which is designed to scale to millions of vertices and edges. Nevertheless, JGraphT relies upon dynamically sized collections, and an array based graph implementation can greatly reduce memory utilization at the cost of additional overhead to keep the data graph consistent with the underlying database
Our experimental results do not reflect well on existing relational keyword search techniques. Runtime performance is unacceptable for most search techniques. Memory consumption is also excessive for many search techniques. Our experimental results question the scalability and improvements claimed by previous evaluations. These conclusions are consistent with previous evaluations that demonstrate the poor runtime performance of existing search techniques as a prelude to a newly-proposed approach.
Many evaluations are also contradictory, for the reported performance of each system varies greatly between

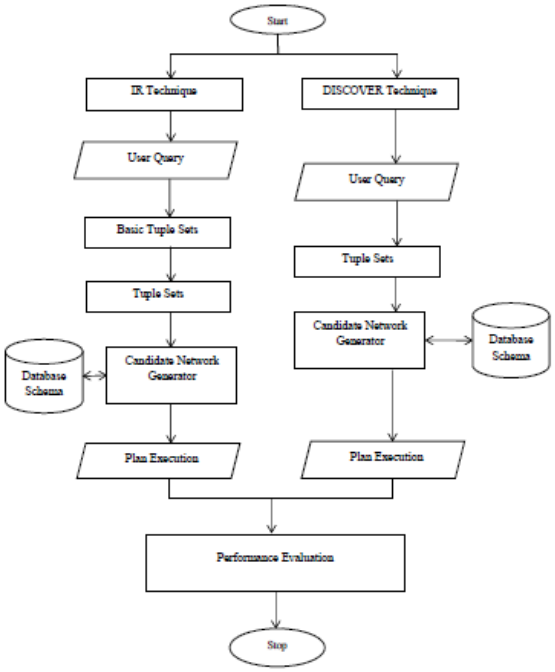
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

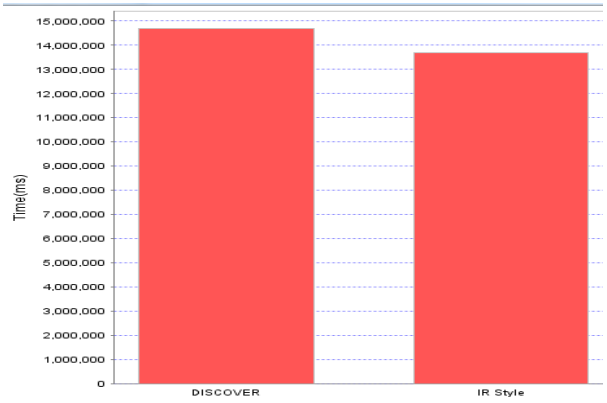
different evaluations. For the offline approaches, the size of the index can exceed the size of the original database by an order of magnitude. Our experimental results question the validity of many previous evaluations, and we believe our benchmark is more robust and realistic with regards to the retrieval tasks than the workloads used in other evaluations.

System flow:

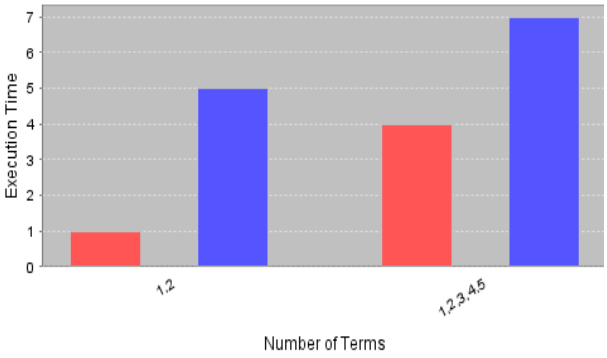


V. RESULT ANALYSIS

Unlike many evaluations reported in the literature, ours investigates the overall, end-to-end performance of relational keyword search techniques.
Hence, we favor a realistic query workload instead of a larger workload with queries that are unlikely to be representative (e.g., queries created by randomly selecting terms from the data set).



Our experimental results do not reect well on existing relational keyword search techniques. Runtime performance is unacceptable for most search techniques.



Memory consumption is not excessive for this search technique. Our experimental results question the scalability and improvements claimed by previous evaluations.

sr. no	Title	Techniques used	Advantages	Disadvantages
1	What Are We Searching For? Analyzing User Objectives When Searching Relational Data. (2012). AUTHOR NAME: J. Coffman and A.C. Weaver	It investigate the types of queries that users might submit to a relational keyword search system. Develop a framework for analyzing queries in this context, which is similar to existing taxonomies for web searches.	1.Improve evaluation of search techniques. 2.Synthetic workloads more representative than existing Workloads. 3.Queries have similar intent and expression.	1.More complex search tasks than are present in existing Internet search engine logs. 2.Keyword search is a simple and flexible alternative,with minimal loss of querying power.
2	Providing Built-in Keyword Search Capabilities in RDBMS (2011) AUTHOR NAME: G. Li, J. Feng, X. Zhou, and J. Wang	it propose a new concept called Compact Steiner Tree , which can be used to approximate the Steiner tree problem for answering top-k keyword quries efficiently.	1.effective, 2.puts the adversary at risk of being detected. 3. increasing the effort in recovering plaintext passwords from the hashed Lists.	such a list may also be available to the adversary, who could use it to help identify oneywords.[2]

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

3	Evaluating the Effectiveness of Keyword Search AUTHOR NAME:W. Webber	the evolving practices and resources for effectiveness evaluation of keyword searches on relational databases	1.Effectiveness of a response to a keyword query. 2. Keyword search offers a straightforward, intuitive, and flexible method of retrieving information.	1.The fundamental technical and formal problems of performing such search. 2. Unbounded in the presence of cycles in the schema graph, and bounded by the size of the data.
4	Ranking Support for Keyword Search on Structured Data Using Relevance Models(2011) AUTHOR NAME: V. Bicer, T. Tran, and R. Nedkov	It Investigate novel strategies for smoothing probabilities in this structured data setting.	1.Results are very encouraging 2. state-of-the-art approaches for the efficient computation of keyword search results.	Finding and ranking relevant resources is the core problem in IR.
5	Toward Scalable Keyword Search over Relational Data (2010). AUTHOR NAME: A. Baid, I. Rae, J. Li, A. Doan, and J. Naughton.	We explore complex relationships among protection techniques ranging from denial and isolation, to degradation and obfuscation, through negative information and deception, ending with adversary attribution and counter-operations.	Outlined a new classification scheme for deception techniques in cyber security	Have shown how some of these techniques have been known and used for many years, but that the field is under-developed

VI. CONCLUSION

Hence we conclude that proposed technique is satisfying number of requirement of keyword query search using different algorithms. The performance of keyword search is also the better to compare other and it shows the actual result rather than tentative. Proposed system investigates the overall end to end performance of relational keyword search techniques which is not obtain in many evaluations reported in literature. It also shows the ranking of keyword and not requires the knowledge of database queries. Compare to existing algorithm it is a fast process.

REFERENCES

[1] Joel Coffman, Alfred C. Weaver,“ An Empirical Performance Evaluation of Relational Keyword Search Systems ”,2014.
[2] J. Coffman and A.C. Weaver, “What Are We Searching For? Analyzing User Objectives When Searching Relational Data,”Proc. Workshop Web Search Click Data (WSCD ’12), Feb. 2012.
[3] G. Li, J. Feng, X. Zhou, and J. Wang, “Providing Built-in Keyword Search Capabilities in RDBMS,” The VLDB J., vol. 20, pp. 1-19, Feb.2011.
[4] W. Webber, “Evaluating the Effectiveness of Keyword Search,”IEEE Data Eng. Bull., vol. 33, no. 1, pp. 54-59, Mar. 2010.
[5] A. Baid, I. Rae, J. Li, A. Doan, and J. Naughton, “Toward Scalable Keyword Search over Relational Data,” Proc. VLDB Endowment, vol. 3, no. 1, pp. 140-149, 2010.
[6] V. Bicer, T. Tran, and R. Nedkov, “Ranking Support for Keyword Search on Structured Data Using Relevance Models,” Proc. 20th ACM Int’l Conf. Information and Knowledge Management (CIKM ’11), pp. 1669-1678, 2011.

BIOGRAPHY

Ms. Madhuri V. Diwan , is currently pursuing M.E (Computer) from Department of Computer Engineering, Dnyanganga College of Engineering and Research, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. She received her B.E (Computer) Degree from S.B. Patil college of engineering, Indapur, Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is information Retrieval and data mining.