ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

Relational Collaborative Data Keywords Search by Empirical Models

Vedakumar.M, P.Nageswara Rao, Bullarao Domathoti

M. Tech Student , Dept. of CSE, Swetha Institute of Technology & Science, JNTUA, Tirupati, India

Associate Professor, Dept. of CSE., SITS, JNT University, Anantapuram, AP, India

Assistant Professor, Dept. of CSE., SITS, JNT University, Anantapuram, AP, India

ABSTRACT: Extending the keys search paradigm to relational data has been an active area of survey with in the database and uses insights from transaction cost economics and agency theory to posit that uncertainty in inter-firm relations increases the difficulty in measuring contractual performance thereby leading to contractual incompleteness. To protect from the resultant contractual opportunism, firms are more likely to use collaborative contracting. In this paper, we present the most extensive empirical performance evaluation of relational keyword search techniques to appear to date in the literature. Our results indicate that many existing search techniques do not provide acceptable performance for realistic retrieval tasks. In particular, memory consumption precludes many search techniques from scaling beyond small data sets with tens of thousands of vertices. We also explore the relationship between execution time and factors varied in previous evaluations; our analysis indicates that most of these factors have relatively little impact on performance. In summary, our work confirms previous claims regarding the unacceptable performance of these search techniques and underscores the need for standardization in evaluation.

KEYWORDS: **Collaboration data; IR Community; keywords;**

I. INTRODUCTION

The ubiquitous search text box has transformed the way people interact with information. Nearly half of all Internet users use a search engine daily [10], performing in excess of 4 billion searches [11]. 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.

However, we are not aware of any research projects that have transitioned from proof-of-concept implementations to deployed systems. We posit that the existing, ad hoc evaluations performed by researchers are not indicative of these systems' real-world performance, a claim that has surfaced recently in the literature [1], [5], [33]. Despite the significant number of research papers being published in this area, existing empirical evaluations ignore or only partially address many important issues related to search performance. Baid et al. [1] assert that existing systems have unpredictable performance, which undermines their usefulness for real-world retrieval tasks. This claim has little support in the existing literature, but the failure for these systems to gain a foothold implies that robust, independent evaluation is necessary. In part, existing performance problems may be obscured by experimental design decisions such as the choice of datasets or the construction of query workloads. Consequently, we conduct an independent, empirical evaluation of existing relational keyword search techniques using a publicly available benchmark to ascertain their real-world performance for realistic query workloads.

A. *Overview of Relational Keyword Search*

Keyword search on semi-structured data (e.g., XML) and relational data differs considerably from traditional IR.1 A discrepancy exists between the data's physical storage and a logical view of the information. Relational databases are normalized to eliminate redundancy, and foreign keys identify related information. Search queries frequently cross

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

these relationships (i.e., a subset of search terms is present in one tuple and the remaining terms are found in related tuples), which forces relational keyword search systems to recover a logical view of the information. The implicit assumption of keyword search—that is, the search terms are related complicates the search process because typically there are many possible relationships between two search terms. It is almost always 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 coverage of search results.

II. LITERATURE SURVEY

1. A Framework for Evaluating Database Keyword Search Strategies:

With regard to keyword search systems for structured data, research during the past decade has largely focused on performance. Researchers have validated their work using ad hoc experiments that may not reflect real-world workloads. We illustrate the wide deviation in existing evaluations and present an evaluation framework designed to validate the next decade of research in this field. Our comparison of 9 state-of-the-art keyword search systems contradicts the retrieval effectiveness purported by existing evaluations and reinforces the need for standardized evaluation. Our results also suggest that there remains considerable room for improvement in this field. We found that many techniques cannot scale to even moderately-sized datasets that contain roughly a million tuples. Given that existing databases are considerably larger than this threshold, our results motivate the creation of new algorithms and indexing techniques that scale to meet both current and future workloads.

2. Keyword Search on Structured and Semi-Structured Data:

Empowering users to access databases using simple keywords can relieve the users from the steep learning curve of mastering a structured query language and understanding complex and possibly fast evolving data schemas. In this tutorial, we give an overview of the state-of-the-art techniques for supporting keyword search on structured and semi-structured data, including query result definition, ranking functions, result generation and top-k query processing, snippet generation, result clustering, query cleaning, performance optimization, and search quality evaluation. Various data models will be discussed, including relational data, XML data, graph-structured data, data streams, and workflows. We also discuss applications that are built upon keyword search, such as keyword based database selection, query generation, and analytical processing. Finally we identify the challenges and opportunities of future research to advance the field.

3. Toward Scalable Keyword Search over Relational Data:

Keyword search (KWS) over relational databases has recently received significant attention. Many solutions and many prototypes have been developed. This task requires addressing many issues, including robustness, accuracy, reliability, and privacy. An emerging issue, however, appears to be performance related: current KWS systems have unpredictable running times. In particular, for certain queries it takes too long to produce answers, and for others the system may even fail to return (e.g., after exhausting memory). 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.

4. Indexing Relational Database Content Offline for Efficient Keyword-Based Search:

Information Retrieval systems such as web search engines offer convenient keyword-based search interfaces. In contrast, relational database systems require the user to learn SQL and to know the schema of the underlying data even to pose simple searches. We propose an architecture that supports highly efficient keyword-based search over relational databases: A relational database is "crawled" in advance, text-indexing virtual documents that correspond to interconnected database content. At query time, the text index supports keyword-based searches with instantaneous response, identifying database objects corresponding to the virtual documents matching the query. Our system, EKSO,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

creates virtual documents from joining relational tuples and uses the DB2 Net Search Extender for indexing and keyword-search processing. Experimental results show that index size is manageable, query response time is indeed instantaneous, and database updates (which are propagated incrementally as recomputed virtual documents to the text index) do not significantly hinder query performance. We also present a user study confirming the superiority of keyword-based search over SQL for a wide range of database retrieval tasks.

5. Bidirectional Expansion for Keyword Search on Graph Databases:

Relational, XML and HTML data can be represented as graphs with entities as nodes and relationships as edges. Text is associated with nodes and possibly edges. Keyword search on such graphs has received much attention lately. A central problem in this scenario is to efficiently extract from the data graph a small number of the "best" answer trees. A Backward Expanding search, starting at nodes matching keywords and working up toward confluent roots, is commonly used for predominantly text-driven queries. But it can perform poorly if some keywords match many nodes, or some node has very large degree. In this paper we propose a new search algorithm, Bidirectional Search, which improves on Backward Expanding search by allowing forward search from potential roots towards leaves. To exploit this flexibility, we devise a novel search frontier prioritization technique based on spreading activation. We present a performance study on real data, establishing that Bidirectional Search significantly outperforms Backward Expanding search.

In existing system, extending the keyword search paradigm to relational data has been an active area of research within the database and information retrieval (IR) community. A large number of approaches have been proposed and implemented, but despite numerous publications, there remains a severe lack of standardization for system evaluations. This lack of standardization has resulted in contradictory results from Different evaluations and the numerous discrepancies muddle what advantages are proffered by different approaches.

III. PROPOSED SYSTEM

In proposed system, empirical performance evaluation of relational keyword search systems. Our results indicate that many existing search techniques do not provide acceptable performance for realistic retrieval tasks. In particular, memory consumption precludes many search techniques from scaling beyond small datasets with tens of thousands of vertices. We also explore the relationship between execution time and factors varied in previous evaluations; our analysis indicates that these factors have relatively little impact on performance. In summary, our work confirms previous claims regarding the unacceptable performance of these systems and underscores the need for standardization as exemplified by the IR community when evaluating these retrieval systems.

Advantages of proposed system:

- Keyword Search with ranking.
- Execution Time consumption is less.
- File length and Execution time can be seen.
- Ranking can be seen by using chart.

when transactional hazards are high, collaborative relationships are more likely than arm's-length transactions Milgrom and Roberts (1992) define a collaborative contract (which they refer to as a “relational contract”) as one that “does not attempt the impossible task of complete contracting but instead settles for an agreement that frames the relationship” (p. 131, emphasis added) and relies on “unarticulated but (presumably) shared expectations that the parties have concerning the relationship” (p. 132). A collaborative relation entails sharing not only information and resources, but also risks and rewards (Kumar 1996). Indeed, confidence and mutual trust exist between the parties because each expects the other to cooperate (Das and Teng 1998; Holmstrom and Roberts 1998). Thus, trust and the repeated exchange associated with collaboration compensate for the lack of adequate performance measures necessary

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

to enforce contractual provisions. Collaboration, which is also associated with alignment of strategic objectives and temporal horizons, can therefore facilitate contracting by increasing trust between the contracting parties. We extend this area of research by first examining the antecedents of collaboration. Specifically, we explore measurability of contractual performance as a factor that drives whether a buyer-seller relation will be collaborative. We next examine the effect of collaborative contracting on relation-specific investments, i.e., investments in assets that have a low value outside the relationship. Ex-post contractual risks are especially pronounced when a supply relationship entails relation-specific investments and uncertainty is high. One example of such a relation-specific investment is an in-process die used in the automobile industry to shape steel sheets into parts for a specific vehicle (Klein, Crawford, and Alchian 1978). These dies, which require significant capital investments by the parts supplier, have little to no value outside the relationship between the automaker and We empirically examine whether the collaborative nature of the relationship affects the likelihood of the supplier making a relation-specific investment. We predict that, because collaboration helps protect firms from contract incompleteness, collaboration reduces the risk of hold-up by the customer and thereby increases the supplier's willingness to invest in relation-specific assets (Parkhe 1993).

1. Graph-based Systems

The objective of proximity search is to minimize the weight of result trees. This task is a formulation of the group Steiner tree problem [9], which is known to be NP-complete [29]. Graph-based search techniques are more general than schema based approaches, for relational databases, XML, and the Internet can all be modeled as graphs. BANKS [2] enumerates results by searching the graph backwards from vertices that contain query keywords. The backward search heuristic concurrently executes copies of Dijkstra's shortest path algorithm [7], one from each vertex that contains a search term. When a vertex has been labeled with its distance to each search term, that vertex is the root of a directed tree that is a result to the query. BANKS-II [17] augments the backward search heuristic [2] by searching the graph forwards from potential root nodes. This strategy has an advantage when the query contains a common term or when a copy of Dijkstra's shortest path algorithm reaches a vertex with a large number of incoming edges. Spreading activation prioritizes the search but may cause the bidirectional search heuristic to identify shorter paths after creating partial results. When a shorter path is found, the existing results must be updated recursively, which potentially increases the total execution time. Although finding the optimal group Steiner tree is NPcomplete, there are efficient algorithms to find the optimal tree for a fixed number of terminals (i.e., search terms). DPBF [8] is a dynamic programming algorithm for the optimal solution but remains exponential in the number of search terms. The algorithm enumerates additional results in approximate order. He et al. [13] propose a bi-level index to improve the performance of bidirectional search [17]. BLINKS partitions the graph into blocks and constructs a block index and intrablock index. These two indices provide a lower bound on the shortest distance to keywords, which dramatically prunes the search space.STAR [18] is a pseudopolynomial-time algorithm for the Steiner tree problem. It computes an initial solution quickly

TABLE I:CHARACTERISTICS OF THE EVALUATION DATASETS

Dataset	[V]	[E]	[T]
MONDIAL	17	56	12
IMDb	1673	6075	1748
Wikipedia	206	785	750

Legend, all values are in thousands [V] number of nodes (tuples) [E] number of edges in data graph [T] number of unique terms

In some instances however, the degree of uncertainty associated with some parts of the process imposes difficulties in measuring performance. In the case of activities such as innovation and R&D, even simply defining, let alone measuring performance is difficult. Additionally, when tasks are nonseparable, i.e., when each party's contribution to the outcome cannot be easily identified, outcome performance measures cannot be relied upon to assess contractual performance. Measuring a supplier's performance may also be complicated by "systems effects," in which the performance of one supplier depends to some extent on either the performance of another supplier or of the customer itself. Thus, identifying performance standards and using those standards to evaluate performance is more difficult for certain services than for others. In such instances, monitoring the supplier's performance and accordingly, determining whether the supplier has fulfilled the terms of the contract becomes more difficult. Under such circumstances,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

contracting between the parties may be hampered by the supplier's unwillingness to risk expending effort without the guarantee of some return. At the same time, the customer may be unwilling to guarantee a return in the absence of measurable outcomes. As a result, formal contractual safeguards cannot be employed.

2. Investments in Relation-Specific Assets Collaborative Contracting:

The risk of opportunistic behavior intensifies in the presence of relation-specific investments when uncertainty is high and contractual outcomes cannot be specified *ex ante*. Williamson (1983) identifies four types of relation-specific investments. These include: (a) site specificity, where the supplier and the customer are located in a "cheek-by-jowl" relation to reduce inventory and transportation expenses, (b) physical asset specificity, where one party or the other must invest in an asset that has no (or significantly less) value outside the relationship for which the investment was made dedicated assets committed to a particular supply arrangement that, if terminated, would leave the firm with considerable excess capacity.

To protect themselves from opportunism when investments in relation-specific assets are made, contracting parties will attempt to employ various safeguards including formal contracts (Dyer 1997). A comprehensive contract that stipulates the obligations and expected actions of each party, as well as the ramifications in the event of unexpected environmental conditions, decreases the risks that a supplier would be exposed to from the relation-specific investments. Indeed, from an agency perspective, when complete contracting is feasible, asset ownership is irrelevant because the contract can assign rights associated with asset ownership (Baiman and Rajan 2002).

However, TCE argues that as the complexity of the contract increases, so does the cost of contracting for both parties. In extreme cases, contracting costs can increase to such a degree that they become prohibitive, requiring the contracting parties to explore other options for safeguarding their relation-specific investments (Dyer 1997). As a result, contracting parties need to establish governance procedures whose safeguards against opportunism are sufficient to increase the supplier's willingness to invest in relation-specific assets. While both the supplier and customer stand to gain from the supplier's investments in relation-specific assets, the distribution of risk thereafter is uneven (recall that once the supplier invests in the relation-specific asset, it exposes itself to the risk of hold-up by the customer).

3. Empirical Models:

Discussions with EMS management and our own examination of the contracts suggest that specifying and measuring the contractual performance of supply chain management services is more complex for several reasons. First, supply chain management services have a longer lead-time. In other words, effort expended at a particular point in time is associated with returns at a different point in time. Second, the payoffs for increased effort in one part of the value chain may accrue at a different part of the value chain. Third, the range of acceptable outcomes is less clear because the activity is inherently more ambiguous. Fourth, supply chain management requires coordination from people in different parts of the supply chain as well as in different functional areas such as engineering, marketing, and finance, each of which may have different measures of performance. Finally, since the outcome of services such as supply chain management depends on the performance of numerous suppliers as well as on the customer, attributing responsibility for the outcome can prove difficult.

Hence, we expect performance to be easiest to measure with repair services and hardest with supply chain management services. In sum, the arguments presented above suggest that uncertainty and accordingly noise in performance measurement will be greater with supply chain management service than manufacturing, and that consequently because of difficulty in assessing performance, monitoring will be lower.

IV. CONCLUSION

Unlike many of the evaluations reported in the literature, ours is designed to investigate not the underlying algorithms but the overall, end-to-end performance of these retrieval systems. 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 dataset). Memory consumption during a search has not been the focus of any previous evaluation. To the best of our knowledge, only two papers [6], [18] have been published in the literature that make allowances for a data graph that does not fit entirely within main memory. Making the original source code (or a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 7, July 2015

binary distribution that accepts a database URL and query as input) available to other researchers would be ideal and greatly reduce the likelihood that observed differences are implementation artifacts.

REFERENCES

1 A. Baid, I. Rae, J. Li, A. Doan, and J. Naughton, "Toward Scalable Keyword Search over Relational Data," Proceedings of the VLDB Endowment, vol. 3, no. 1, pp. 140–149, 2010.
2 G. Bhalotia, A. Hulgeri, C. Nakhe, S. Chakrabarti, and S. Sudarshan, "Keyword Searching and Browsing in Databases using BANKS," in Proceedings of the 18th International Conference on Data Engineering, ser. ICDE '02, February 2002, pp. 431–440.
3 S. Chaudhuri and G. Das, "Keyword Querying and Ranking in Databases," Proceedings of the VLDB Endowment, vol. 2, pp. 1658–1659, August 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1687553. 1687622
4 Y. Chen, W. Wang, Z. Liu, and X. Lin, "Keyword Search on Structured and Semi-Structured Data," in Proceedings of the 35th SIGMOD International Conference on Management of Data, ser. SIGMOD '09, June 2009, pp. 1005–1010.
5 J. Coffman and A. C. Weaver, "A Framework for Evaluating Database Keyword Search Strategies," in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ser. CIKM '10, October 2010, pp. 729–738. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871531
6 B. B. Dalvi, M. Kshirsagar, and S. Sudarshan, "Keyword Search on External Memory Data Graphs," Proceedings of the VLDB Endowment, vol. 1, no. 1, pp. 1189–1204, 2008.
7 E. W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," Numerische Mathematik, vol. 1, no. 1, pp. 269–271, 1959.
8 B. Ding, J. X. Yu, S. Wang, L. Qin, X. Zhang, and X. Lin, "Finding Topk Min-Cost Connected Trees in Databases," in ICDE '07: Proceedings of the 23rd International Conference on Data Engineering, April 2007, pp. 836–845.
9 S. E. Dreyfus and R. A. Wagner, "The Steiner Problem in Graphs," Networks, vol. 1, no. 3, pp. 195–207, 1971. [Online]. Available: http://dx.doi.org/10.1002/net.3230010302
10 K. Golenberg, B. Kimelfeld, and Y. Sagiv, "Keyword Proximity Search in Complex Data Graphs," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '08, June 2008, pp. 927–940.