Keyword Search and Evaluation over Relational Databases: an Outlook to the Future

Sonia Bergamaschi
University of Modena and Reggio Emilia, Italy
sonia.bergamaschi@unimore.it

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Francesco Guerra
University of Modena and Reggio Emilia, Italy
francesco.guerra@unimore.it

Gianmaria Silvello
University of Padua, Italy
silvello@dei.unipd.it

ABSTRACT

This position paper discusses the need for considering keyword search over relational databases in the light of broader systems, where keyword search is just one of the components and which are aimed at better supporting users in their search tasks. These more complex systems call for appropriate evaluation methodologies which go beyond what is typically done today, i.e. measuring performances of components mostly in isolation or not related to the actual user needs, and, instead, able to consider the system as a whole, its constituent components, and their inter-relations with the ultimate goal of supporting actual user search tasks.

1. INTRODUCTION

Keyword search is the foremost approach for information searching and in the last decades it has been extensively studied in the field of Information Retrieval (IR). Nevertheless, this model has left out the structured data sources which are typically accessed through structured queries. Structured queries are not suitable for the generic user, since their formulation requires users to know a language and the data source schemas and contents. The large availability of structured data has made of paramount importance the development of keyword search tools for this kind of sources.

In relational databases, keyword search aims at easing and somehow automating the search process by employing two main techniques [8]: schema-based and graph-based. Graph-based techniques model relational databases as graphs, where nodes are tuples, edges foreign-primary key relationships between those tuples and the algorithms are based on the computation of specific structures over the graphs. Whereas, schema-based techniques exploit the schema information to formulate Structured Query Language (SQL) queries determined starting from the user keyword queries.

DBRank 2013, August 30 2013, Riva del Garda, Italy

http://dx.doi.org/10.1145/2524828.2524836

On the other hand, it should be considered that search process is part of a wider user task [3] from which the information need arises and, in turn, makes the user resort to issuing queries to satisfy it. This makes the whole process quite complex and brings in the accomplishment of the user information need several degrees of uncertainty.

Even if the main focus of keyword search, i.e. getting out the most from the relational data starting from a keyword instead of a structured query, is certainly something that helps users in carrying out their tasks, many factors, beyond algorithmic correctness and completeness, may impair the impact of keyword search and prevent users to fully exploit its potentialities. Therefore, we need to put keyword search into a broader context and envision innovative architectures where keyword search is one of the components, paired with other building blocks to better take into account the variability and uncertainty entailed by the whole search process.

We claim that a conceptual architecture pivoting around keyword search and relational data needs to couple system- and user-oriented components; the former ones aim at augmenting the performances (i.e. efficiency) of the search, whereas the latter ones aim at improving the quality of the search (i.e.effectiveness) from the user perspective. Such system- and user-oriented components already exist and have been demonstrated to be effective for their specific purpose. Nevertheless, their integration into a unique framework for keyword search is still lacking. Evidence of this is the absence of a commercially deployed system.

Moreover, measuring is a key to scientific progress and experimental evaluation – both laboratory and interactive – is a key means for supporting and fostering the development of systems which are more adherent to the user needs, provide the desired effectiveness and efficiency, guarantee the required robustness and reliability, and operate with the necessary scalability. In light of this, we claim that the current frameworks for the evaluation of keyword search in relational databases [1] need to be re-thought, by moving beyond the evaluation of keyword search components in isolation or not related to the actual user needs, and, instead, by considering the whole system, its constituents, and their inter-relations with the ultimate goal of supporting actual user search tasks.

The paper is organized as follows: Section 2 describes the main components of the proposed approach; Section 3 de-

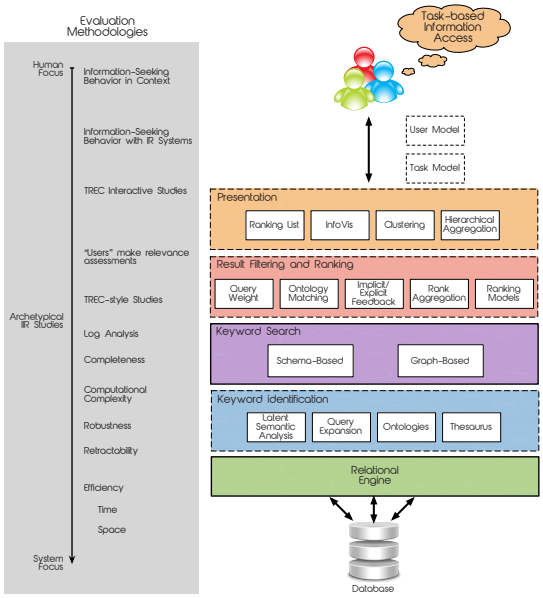

Figure 1: The Conceptual Architecture of a Keyword-based Search System and Evaluation.

fines the main characteristics of a framework for evaluating such an innovative keyword search system. Finally, Section 4 draws some final remarks.

2. CONCEPTUAL ARCHITECTURE

In Figure 1 we can see, on the right-hand side, a general architecture of an information access system pivoting on keyword search techniques, and on the left-hand side, the evaluation methodologies that can be employed for evaluating the efficiency and efficacy of the system. Thick solid lines frame modules which are the focus of current keyword search systems, whereas dotted lines frame components which are typically not exploited today and come from neighboring fields, such as information retrieval, information extraction, data mining, and natural language processing.

The main keyword search and relational database layers are surrounded by a Keyword Identification layer, a Results Filtering and Ranking layer, and a Presentation layer. Indeed, an ideal search system has to consider the search task a user desires to conduct, to perform user queries knowing that they may not exactly correspond to the real user information need, to disambiguate search terms, to rank the results of the search process on the basis of relevance for the user, and to visualize these results in the most proper way for the considered search task.

The *Keyword Identification* layer is aimed at freeing keyword search from an exact match with the keywords present in the relational data and introduce the possibility of matching in multiple ways the keywords expressed by user to the relational data in order to compensate for possible imprecisions or errors in the choice of the keywords.

The *Results Filtering and Ranking* layer accounts for the need of adopting alternative strategies for ranking and selecting the results to be presented to the user by the system. It may concern weighting the results on the basis of the process which generated relational queries from user keywords, or relying on implicit/explicit feedback from the user to filter out some results, or using rank aggregation and data fusion techniques [7] to merge alternative ranking strategies.

The *Presentation* layer regards how the outputs of a system are presented to the user; for instance, we can have traditional ranking lists, results presentation based on advanced Information Visualization techniques [9], or presentations of clusters of results.

3. EVALUATION

Innovative proposals for pushing the boundaries of keyword search cannot set aside a proper and shared evaluation methodology which helps in progressing towards the envisioned goals, ensures the soundness and quality of the proposed solutions, and guarantees the repeatability and comparability of the experiments.

As shown in Figure 1, evaluation can be carried out at three levels: at a "task level" for instance by means of user studies [4]; at an "effectiveness level" by means of the test collection methodology [2]; and, at a "system level" by means of benchmarking queries per second, memory and CPU load, correctness and completeness [1], thus moving from a human to a system focus.

Nevertheless, experimental evaluation is hampered by fragmentation – different tasks, different collections, different perspectives from interactive to laboratory evaluation which are usually dealt with in separated ways, without sharing resources and the produced data [6]. This will be even more true for the multidisciplinary approach to the system architecture proposed on the right of Figure 1, whereas the unified and holistic approach to evaluation, proposed on the left, would be needed to assess the different facets of such a complex system and to reconcile the experimental outcomes.

Furthermore, systematic and comparable experimental evaluation is a very demanding activity, in terms of both time and effort needed to perform it. For this reason, in the IR field, it is usually carried out in publicly open and large-scale evaluation campaigns at international level, which allow for sharing the effort, producing large experimental col-

lections, and comparing state-of-the-art systems and algorithms. Relevant and long-lived examples are TREC (http://trec.nist.gov/) in the United States, the CLEF initiative (http://www.clef-initiative.eu/) in Europe, and NTCIR (http://research.nii.ac.jp/ntcir/) in Japan and Asia. Relying on these experiences would move evaluation of keyword search forward, by sharing resources, providing open fora to compare and discuss approaches, re-use the data and the acquired knowledge, have the possibility of repeating the obtained results, and reduce the overall effort.

Moreover, as reported by [5], for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to researchers and industry. During their lifespan, large-scale evaluation campaigns have produced huge amounts of scientific data which are extremely valuable for research and development but also from an economic point of view: [5] estimates that the overall investment in TREC of about 30 million dollars in its first 20 years which, as discussed above, produced an estimated return on investment between 90 and 150 million dollars. Therefore, applying experimental evaluation to vision represented in Figure 1 gives the promise not only to advance state-of-the-art techniques, but also to have a concrete economic impact.

4. USE CASE

Let us consider a relational database containing tourism information about accommodations. For sake of simplicity, the database is composed of two tables: Accommodation and City, describing respectively accommodations and the cities where the accommodations are located. Some descriptive attributes are defined for each table and a foreign key - primary key relationship exists between the attribute `City` in table Accommodation and the attribute `Name` in table City.

By traditional keyword search techniques, even the simple query "Hotel Venice" may not be trivial to answer. Firstly, an index-based approach can find several matches for the keyword "Venice" with instances of the table Accommodation (i.e., Venice is the `Name` of an accommodation, or Venice is a value of the attribute `City` in the same table) or with an instance of the table City (i.e Venice is the name of the city). This is due to the fact that the user query is ambiguous and it is unclear whether the user is looking for hotels called "Venice" or hotels which are in "Venice". Moreover, a classical keyword search engine may not find any correspondences between the keyword "Hotel" and an element in the database. In this case, the problem is two-fold: firstly, the chosen keyword may better match metadata instead of a value in a table – indeed, several existing keyword search engines do not consider database metadata as possible targets for user queries; secondly, the chosen keyword may actually refer to a concept represented in the table with a value which is a synonym of the chosen keyword. Since users do not know the information in the data source, this situation frequently occurs and is typically not managed by most of the existing systems.

This problem can be even more complex if some keywords match with foreign key - primary key values. This occurs in our example if the user formulates the query "Venice", where it is unspecified if the user is interested in accommodations or cities. From an algorithmic perspective an answer showing all the hotels in Venice provides an answer as good as the one that reports all the information about the city. Nevertheless, from the user perspective, only one of these two sets of results is relevant to her/his information needs, e.g. the city one, while providing both of them would reduce the effectiveness of the system and hampers performance.

Even in this toy example, it becomes clear that query ambiguity as well as the choice between graph or schema based techniques impacts the system performance. Therefore, the complexity of real scenarios may take even more advantage from the architecture proposed in Figure 1, which complements keyword search with additional components that, in this specific example, analyze the user keywords and disambiguate their meaning. Similarly, the evaluation methodology must be able to detect these different issues in order to properly assess how systems tackle them.

As a further example, it is not clear from the query which is the information that the user would like to receive as a result. Several options are possible: in some cases the user would like to receive all the data about the tuples satisfying the criteria defined by the keyword query (i.e., the "universal relation"), in other cases only the values of some attributes (i.e., a projection of the "universal relation"), or even a boolean value (i.e., the existence of at least an instance). On the other hand, she/he also would like to receive results ordered by the system estimation of their relevance and how much they satisfy her/his information need. The task of understanding the granularity, the ordering and aggregation of the expected results is not typically managed by the existing systems and requires a specific module in the architecture of a complete keyword search system. As above, evaluation must be able to take into account these aspects and assess the systems accordingly.

Acknowledgements

The reported work has been partially supported by PROMISE FP7 network of excellence (http://www.promise-noe.eu/, contract n. 258191) and by the KEYSTONE (http://www.keystone-cost.eu/keystone/, IC1302) cost action.

5. REFERENCES

[1] J. Coffman and A. C. Weaver. An Empirical Performance Evaluation of Relational Keyword Search Techniques. *IEEE TKDE*, 2013 pre-print.

[2] D. K. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA, 2011.

[3] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer, Heidelberg, Germany, 2005.

[4] D. Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.*, 3(1–2):1–224, Jan. 2009.

[5] B. R. Rowe, D. W. Wood, A. L. Link, and D. A. Simoni. *Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program*, July 2010.

[6] G. Weikum. Where's the Data in the Big Data Wave? ACM SIGMOD Blog, March 2013.

[7] S. Wu. *Data Fusion in Information Retrieval*. Springer-Verlag, Heidelberg, Germany, 2012.

[8] J. X. Yu, L. Qin, and L. Chang. Keyword Search in Relational Databases: A Survey. *IEEE Data Eng. Bull.*, 33(1):67–78, 2010.

[9] J. Zhang. *Visualization for Information Retrieval*. Springer-Verlag, Heidelberg, Germany, 2008.