

Searching, Analyzing and Exploring Databases

Yi Chen¹, Wei Wang², and Ziyang Liu¹

¹ Arizona State University

{yi, ziyang.liu}@asu.edu

² University of New South Wales, Australia

weiw@cse.unsw.edu.au

1 Introduction

Keyword based search, analysis and exploration enables users to easily access databases without the need to learn a structured query language and to study possibly complex data schemas. Supporting keyword based search, analysis and exploration on databases has become an emerging hot area in database research and development due to its substantial benefit. Researchers from different disciplines are working together to tackle various challenges in this area.

This tutorial aims at outlining the problem space of supporting keyword based search, analysis and exploration on databases, introducing representative and state-of-the-art techniques that address different aspects of the problem, and discussing further challenges and potential future research directions. The tutorial will provide the researchers and developers a systematic and organized view on the techniques related to this topic.

A tutorial with similar topic was given in SIGMOD 2009 [1]¹, and was very well received. Since the research interest in keyword search on structured data is ever increasing and there are plenty of new techniques since then, this tutorial will be updated to incorporate the new findings in this area, which covers query processing, type ahead search, query suggestion, personalization, result comparison, faceted search, etc.

2 Tutorial Outline

2.1 Search Result Definition, Generation, Ranking and Evaluation

The first task in keyword search is to define query results which *automatically* gather *relevant* information that is generally fragmented and scattered across multiple places.

Query Result Definition. A query result on a graph data model is commonly defined as a subtree of the data graph where no node or edge can be removed without losing connectivity or keywords contained in the subtree. Since finding the smallest result, which is the group Steiner tree, is NP-hard, variations and relaxation of the definition have been proposed in order to attain reasonable efficiency. For example, when the data is modeled as a tree, lowest common ancestor (LCA) is a common form to define the query results. Furthermore, besides the data that match query keywords, studies have been performed on identifying data that do not match keywords or on the paths connecting keyword nodes, but are *implicitly* relevant.

¹ http://www.public.asu.edu/~ychen127/keyword_sigmod09_tutorial.pptx

Ranking Functions. It is almost always the case that not all results of a keyword search are equally relevant to a user. Various ranking schemes have been proposed to rank the results so that users can focus on the top ones, which are hopefully the most relevant ones. Many ranking schemes are used in existing works, which consider both the properties of data nodes (e.g., TF*IDF, node weight, and page-rank style ranking, etc.) and the properties of the whole query result (e.g., number of edges, weights on edges, size normalization, redundancy penalty, etc.).

Result Generation and Top- k Query Processing. Algorithms for query result generation and efficient top- k query processing have been developed, and will be introduced in this tutorial. For example, encoding, indexing schemes as well as materialized views have been exploited for processing keyword search on XML. For keyword search on relational databases and graphs, many approaches are based on candidate network (CN) generation, and there are also dynamic programming, heuristics-based approaches (e.g., backward exploration), indexed search approaches, etc., for generating top- k query results.

Evaluation. We will present and discuss evaluation frameworks for keyword search engines. One type of evaluation framework is based on empirical evaluation using benchmark data, such as INEX (INitiative for the Evaluation of XML Retrieval), a benchmark for XML keyword search. Another type of evaluation is evaluating an approach based on a set of axioms that capture broad intuitions.

2.2 Query Suggestion and Result Analysis

To improve search quality and users' search experience, various techniques have been proposed.

Result Snippets. Result snippets should be generated by most Web search engines to compensate the inaccuracy of ranking functions. Generating snippets for structured search results have also been studied. The principle of result snippets is orthogonal to that of ranking functions: letting the user quickly judge the relevance of query results by providing a brief quotable passage of each query result.

Result Clustering and Comparison. Since many keyword queries are ambiguous, it is often desirable to cluster query results based on their similarity, so that the user can quickly browse all possible result types choose the sets of results that are relevant. Besides, techniques have been developed to automatically generate comparison tables for user selected results to help users analyze and differentiate results and get useful insight from the comparison.

Query Cleaning and Suggestion. User issued keyword queries may not be precise. Query cleaning involves semantic linkage and spelling corrections of query keywords, as well as segmentation of nearby query keywords so that each segment corresponds to a high quality data term. Query suggestion is another way of helping user issue queries, which is through suggesting related queries given the query initially submitted by the user. Query suggestion when searching structured data has recently been studied.

2.3 Other Exploration Methods

Query Form Based Database Access. Since structured query languages are highly expressive but difficult to learn, and keyword queries are easy to use but lack the expressive power, a natural idea is to strike a good balance between the two. Existing attempts include generating a large set of query forms and selecting the relevant forms based on user's keyword query, or directly generating a small set of query forms based on either a sample query workload or properties of the schema and data, etc.

Faceted Navigation. A faceted navigation approach generates a navigation tree, either based on the user's keyword query or directly for the entire data collection, in which each level is a classification of the query result. It helps user narrow down the browse scope and find the relevant results quickly. The main challenge is to select the optimal classification at each level of the tree to minimize the expected navigation cost.

2.4 Open Challenges and Future Directions

We will discuss open problems and possible directions for future research. For example, there are many other types of structured data besides normal trees and graphs, whose structures can be exploited to provide high-quality search results. There are many opportunities for supporting keyword search on data models like data warehouses, spatial and multimedia databases, workflows, and probabilistic databases, as well as data extracted from text documents (e.g. parse tree databases). Furthermore, techniques that enable users to seamlessly access vast collections of heterogeneous data sources are also in demand.

Besides, few existing work on searching structured data connects the quality of search results with user needs. In this aspect, there are much we can learn from the Information Retrieval field. In particular, having user involvement during the search process, such as analysis of query log and user click-through streams, will be helpful to provide personalized search experience. Nonetheless keyword search on structured data poses unique challenges on analyzing user preferences.

To summarize, this tutorial presents the state-of-the-art approaches for keyword based search, analysis and exploration on databases with an emphasis on introducing the different modules of a keyword search engine, the research challenges and the state-of-the-art of each module, as well as their relationships. We hope this tutorial will effectively help the audience get a big picture of this research topic.

Acknowledgement

Yi Chen is supported by NSF CAREER award IIS- 0845647 and NSF grant IIS-0740129, IIS-0915438 and IBM faculty award. Wei Wang is supported by ARC Discovery Grants DP0987273 and DP0881779.

References

1. Chen, Y., Wang, W., Liu, Z., Lin, X.: Keyword Search on Structured and Semi-Structured Data. In: SIGMOD Conference, pp. 1005–1010 (2009)