Efficient Top-k Keyword Search Over Multidimensional Databases

Ziqiang Yu, School of Computer Science and Technology, Shandong University, Jinan, China
Xiaohui Yu, School of Computer Science and Technology, Shandong University, Jinan, China & School of Information Technology, York University, Toronto, Canada
Yang Liu, School of Computer Science and Technology, Shandong University, Jinan, China

ABSTRACT

Keyword search over databases has recently received significant attention. Many solutions and prototypes have been developed. However, due to large memory consumption requirements and unpredictable running time, most of them cannot be applied directly to the situations where memory is limited and quick response is required, such as when performing keyword search over multidimensional databases in mobile devices as part of the OLAP functionalities. In this paper, the authors attack the keyword search problem from a new perspective, and propose a cascading top-k keyword search algorithm, which generates supernodes by a branch and bound method in each step of search instead of computing the Steiner trees as done in many existing approaches. This new algorithm consumes less memory and significantly reduces the response time. Experiments show that the method can achieve high search efficiency compared with the state-of-the-art approaches.

Keywords: Branch and Bound, Keyword Search, Multidimensional Database, Ranking, Supernode

1. INTRODUCTION

The success of search engines on the World Wide Web has popularized the use of keyword search, with which even untrained users are able to find information of interest from vast collections of documents/Webpages. Over the past decade, this success has inspired much interest in studying keyword search over databases, which promises to allow users lacking knowledge of structured query languages or unaware of the database schema to query the database in an intuitive way.

Existing approaches to keyword search over databases can be broadly classified into two categories: those based on the schema graph (Agrawal, Chaudhuri, & Das, 2002), and those based on the data graph (Aditya et al., 2002). Data graph-based approaches, the focus of this paper, model the database as a graph, where nodes correspond to tuples and edges correspond...
to primary-foreign-key relationships. In most of those approaches, the problem of searching for results can be reduced to finding the Steiner trees (Li, Zhou, Feng, & Wang, 2009) that cover all or some of keywords, which is known to be NP-hard (Bhalotia, Hulgeri, Nakhe, Chakrabarti, & Sudarshan, 2002).

In this paper, we consider the problem of providing keyword search capabilities over multidimensional databases where the size of main memory is very limited, such as in mobile devices. In a multidimensional database, the values of any attribute and the cell values in the cubes can all be viewed as data nodes. Furthermore, we can regard the relationships among the cube values and the corresponding attribute values as edges. Thus, a multidimensional database can be viewed as a data graph. Performing keyword search on multidimensional databases is therefore essentially equivalent to performing the search in data graphs.

Mobile devices, such as smart phones and PDAs, have gained huge popularity over the past few years and are now used extensively for business purposes. It is therefore highly desirable to provide OLAP functionalities in those devices. Some studies have already started to address this problem (Maniatis, 2004; Cuzzocrea & Furfaro, 2003), where OLAP operations are performed over data stored in the devices. Such data are often a selection of the data from the backend server, sometimes appearing in transformed forms, e.g., multidimensional databases (Cuzzocrea, Furfaro, & Sacca, 2009; Liu, Lee, & Lee, 2005). However, there has been no prior work on keyword search in such contexts, although keyword search is a particularly good interface for OLAP (and business intelligence in general) applications, primarily due to its ease of use.

One major difficulty in deploying existing keyword search approaches to mobile devices is that they do not pay attention to the characteristics of mobile devices, such as small memory size and limited power. Schema graph-based methods generate SQL statements to carry out the search. This often involves a lot of I/Os and can be a power hog. Data graph-based methods, on the other hand, typically assume that all required data structures can fit in main memory, which can be unrealistic in the mobile context. Although Dalvi, Kshirsagar, and Sudarshan, (2008) propose a method to overcome the limit of memory when the data graph are over-large, it need to frequently transfer a amount of data from disks to RAM, which greatly exceeds the processing power of mobile devices. Moreover, most of the existing algorithms consider search effectiveness a top priority and are optimized for that, but they may take a long (and unpredictable) time to run, which are not suitable for mobile applications where swift responses are normally expected. Finally, many keyword search algorithms are so complex that they go beyond the weak processing power of the mobile devices. Such observations motivate us to develop novel keyword search techniques that have small memory consumption requirements with fast response times.

We propose a Cascading Top-k Keyword Search algorithm (CTKS for short), a new method to produce answers to keyword search queries quickly in a space-efficient way. The basic idea is to process the keywords in the query sequentially according to their relative importance; each step of the search then corresponds to finding the best match for the next yet-to-be-processed keyword. To further reduce the complexity of the search, each newly found matching node is merged with the existing matching nodes (and any nodes along the path) into a super-node, which allows us to convert the problem of searching for the next keyword into one of finding the shortest path between the supernode and the node containing the next keyword. We further extend this scheme by keeping the top-k results at each step, where k can be adjusted to reflect the memory consumption requirements of the applications. In addition, we introduce a ingenious branch and bound method to optimize generating top-k supernodes at each step, which can greatly improve the efficiency of CTKS. These characteristics of the CTKS algorithm make it well suited for deployment in mobile devices to enrich users’ OLAP experience.
TBSGM: A Fast Subgraph Matching Method on Large Scale Graphs
[www.igi-global.com/article/tbsgm/215006?camid=4v1a](www.igi-global.com/article/tbsgm/215006?camid=4v1a)

Discovery of Emergent Sorting Behavior using Swarm Intelligence and Grid-Enabled Genetic Algorithms
Dimitris Kalles, Alexis Kaporis, Vassiliki Mperoukli and Anthony Chatzinouskas (2014). *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining* (pp. 249-271).