Answering Top-k Keyword Queries on Relational Databases

Myint Myint Thein, University of Computer Studies, Mandalay, Myanmar
Mie Mie Su Thwin, University of Computer Studies, Mandalay, Myanmar

ABSTRACT

Keyword search in relational databases allows the user to search information without knowing database schema and using structural query language. As results needed by user are assembled from connected tuples of multiple relations, ranking keyword queries are needed to retrieve relevant results. For a given keyword query, the authors first generate candidate networks and also produce connected tuple trees according to the generated candidate networks by reducing the size of intermediate joining results. They then model the generated connected tuple trees as a document and evaluate score for each document to estimate its relevance. Finally, the authors retrieve top-k keyword queries by ranking the results. In this paper, the authors propose a new ranking method based on virtual document. They also propose Top-k CTT algorithm by using the frequency threshold value. The experimental results are shown by comparison of the proposed ranking method and the previous ranking methods on IMDB and DBLP datasets.

Keywords: Candidate Network, Connected Tuple Tree, Keyword Query, Keyword Search, Relational Databases, Top-K

1. INTRODUCTION

The most critical and valuable amount of data such as business data has been stored in relational databases. Relational database management system (RDBMS) is a DBMS in which data is saved in tables and the relationships among the data are saved in tables. The data can be reassembled and accessed in many different ways without change the table forms. Most commercial RDBMS uses a structured query language (SQL) to access the database. With more and more data being stored in relational database, it has become crucial for users to be able to search and browse the information stored in them. Keyword search in relational databases enables ordinary users, who do not understand the database schema and SQL, to find the connected tuple sets among the tuples stored in relations, with a given set of keywords. The existing methods of keyword search in relational databases can be broadly classified into two categories that are schema based method and graph based method.

In schema based keyword search in relational database, it has a common method that is generating candidate network in schema graph transformed from relations. Data is stored in

DOI: 10.4018/ijirr.2012070103
the form of columns, tables and primary key to foreign key relationships in relational databases. According to develop the schema graph, we illustrate two schema graphs as examples. Figure 1 shows the schema graph of publication database from DBLP dataset. The movies database schema graph of IMDB dataset shows in Figure 2. For a given keyword query, the logical unit of answers needed by users is not limited to an individual column value or even an individual tuple. It may be multiple tuples joined together. Given keyword search in relational databases, generating minimum joining tuples sets of relations that contained keyword is called candidate network (CN), such as SQL. A candidate network must satisfy the two conditions, total and minimal. Because it is meaningless if two tuples in a candidate network are too far away from each other, the maximum numbers of tuples allowed in a candidate network are needed to specify (Yu & Qin, 2010).

Suppose user wants to get the papers written by “Jinlin Chen” from DBLP database. The system generates the relevant CNs, such as Person ⋈ Relation-Person-InProceeding ⋈ InProceeding, with multiple tuples from different relations joined by foreign keys.

Generating all valid candidate networks that are called connected tuple trees (CTTs) by joining connected tuple trees that can be results for the query. These results are not surely useful to the user. We need to compute a single score for each CN in order to rank the relevant results. So, a ranking method is essential for getting user satisfaction. For the ranking method, some systems considered each text column as a collection and each value in the text column as document by using IR weighting methods. The results are ranked according to a final score that is obtained by dividing the sum of all these scores by the number of tuples in the tuple trees. These methods can help improve the keyword search quality in relational database.

Recently, several researchers have been done on keyword search systems for relational databases (Baid, 2010; Qin & Yu, 2009; Stefanidis & Drosou, 2010; Wang, 2007; Li & Zhu, 2008). In candidate network generation, DISCOVER (Hristidis & Papakonstaninou, 2002) and Liu, Yu, and Meng (2006) developed the CN generation algorithm based on a breadth-first traversal in the search space. SPARK (Luo

**Figure 1.** Publication database schema graph
Hamming Distance based Clustering Algorithm
www.igi-global.com/article/hamming-distance-based-clustering-algorithm/72704?camid=4v1a