Yogesh Bedekar
March 28th, 2008
CSE 718 – Advanced Topics in Database Systems

Effective Keyword Based Selection of Relational Databases

Overview

This paper proposes an effective method to summarize the relationships between keywords in a database based on its structure. It then discusses effective ranking methods based on the keyword relationship summaries in order to select the most useful databases for a given keyword query.

The following are the four main contributions of this paper:

1. Addresses the problem of selection of structured data sources for keyword based queries.
2. Proposes a method for summarizing relationships between keywords in a database.
3. Defines metrics to rank source databases given a keyword query based on keyword relationships.
4. It evaluates the proposed summarization using real datasets.

The paper uses two factors to measure the strength of the relationships between two keywords:

i) A proximity factor which represents the number of joins required to produce a keyword combination.

ii) A frequency factor which indicates the number of ways in which relations can be joined to produce the given keyword combination.

The presence of keywords in various database tuples and the relationships between various tuples are modeled mathematically using matrices. The paper describes an incremental method to compute the shortest join sequence for each pair of keywords which is then used to calculate the frequency factor. All these computations are performed by inserting into and querying from suitably designed database tables.

Having mathematically computed the relationship scores between each pair of keywords, the paper describes four different formulae to determine the relationship score of an entire database over a given query. These four include using the minimum relationship score, the maximum score, sum of the scores and product of the scores over each pair of keywords in the query. On the basis of this score, different databases are ranked for a particular query such that the higher the score of the database, the lower is its rank.

The authors conclude by describing comprehensive experiments performed over multiple databases distributed across a network whose results have been compared with brute force ranking methods. The comparisons between relationship scores determined using the technique described in this paper and the actual brute force evaluation illustrate the efficacy of the technique described in this paper.

Detailed Comments

1. The main strength of this paper is that it proposes a new method to summarize the relationships between the keywords in a database based on its structure. It thus makes a valuable contribution towards selection of relevant data sources in distributed databases. The summarization technique discussed in this paper effectively captures relationships between keywords in a normalized database.
   Traditional IR techniques based on keyword frequency statistics fail to capture these relationships in a normalized database.
   A weakness of this paper is that the technique described in this paper has significant time and storage requirements. For instance, the authors have performed an experiment on a database with 52106 tuples and 20956 distinct keywords. In this experiment, a sharp spike is noted in the time required to compute the summaries when the maximum number of joins to be considered is increased from three to four. The size of the tuple relationship matrices also increases with increase in the number of joins, but the matrices are essentially sparse.

2. The paper is well balanced in terms of technical depth. It begins by clearly describing the motivation behind the proposed technique and a clear description of the basis used for computing keyword relationship summaries. It does discuss many mathematical propositions, but many of these are intuitive in nature and are a logical consequence of the concepts previously discussed.

3. The paper is technically sound. The implementation details described for computing keyword based summaries follow a logical pattern to the concepts discussed earlier. Having discussed the conceptual and mathematical aspects of keyword based summarization in a database, the paper provides the results of experiments performed over distributed databases which demonstrate the effectiveness of the summarization technique described in the paper.

4. Most of the related work in this field focuses on keyword based query processing in a centralized database. This paper proposes an effective method to rank and select useful data sources in a distributed environment for keyword queries, after which centralized keyword based search can be applied on the selected data sources. There have also been several research works on capturing relationships between words in various documents based on word co-occurrence. However such natural language techniques do not address the requirements of a relational database that require us to capture relationships based on tuple references. This paper describes an effective technique for summarization of relational databases.