Chapter 12

Cost Modeling and Range Estimation for Top-\(k\) Retrieval in Relational Databases

Anteneh Ayanso
Brock University, Canada

Paulo Goes
University of Arizona, USA

Kumar Mehta
George Mason University, USA

ABSTRACT

Relational databases have increasingly become the basis for a wide range of applications that require efficient methods for exploratory search and retrieval. Top-\(k\) retrieval addresses this need and involves finding a limited number of records whose attribute values are the closest to those specified in a query. One of the approaches in the recent literature is query-mapping which deals with converting top-\(k\) queries into equivalent range queries that relational database management systems (RDBMSs) normally support. This approach combines the advantages of simplicity as well as practicality by avoiding the need for modifications to the query engine, or specialized data structures and indexing techniques to handle top-\(k\) queries separately. This paper reviews existing query-mapping techniques in the literature and presents a range query estimation method based on cost modeling. Experiments on real world and synthetic data sets show that the cost-based range estimation method performs at least as well as prior methods and avoids the need to calibrate workloads on specific database contents.

INTRODUCTION

In many relational database applications, end-users are more interested in conveniently exploring the data around some attribute values of interest. However, relational queries establish rigid qualification to deal only with data that exactly match selection conditions (Motro, 1988). Due to the exactness in nature of the query model in the RDBMS, users face the challenge of routinely

DOI: 10.4018/978-1-60960-521-6.ch012
specifying value ranges of attributes in search of approximate matches, particularly in the event of empty answers or too many answers (Huh and Lee, 2001; Shin, Huh, Park, and Lee, 2008). Top-k querying enhances the information retrieval (IR) capability of RDBMSs and provides mechanisms for effective retrieval of data that are more appealing to end-user (Ilyas et al., 2006).

Supporting top-k querying in existing RDBMSs, however, is a challenging task and the topic has received significant attention from the database community in recent years (Ilyas, Beskales, and Soliman, 2008). Much of the existing research has focused on techniques that involve high-performance indexing which requires significant changes in the relational query engines. Our technique presented here is motivated by the mechanism of mapping a top-k query into an equivalent range query that is normally supported by RDBMS (Bruno, Chaudhuri, and Gravano, 2002; Chaudhuri and Gravano, 1999; Donjerkovic and Ramakrishnan, 1999). In this approach, top-k querying is defined as a method of specifying relational queries by target values of attributes in order to obtain a desired number of “best matches” based on some ranking (distance) functions. In contrast to using new query operators or specialized index structures to support top-k queries, the query-mapping approach relies on summary statistics that are available in database systems in the form of histograms. Using these statistics, the approach deals with techniques for estimating approximate range queries for efficient top-k retrieval.

Given the enormous installed base of relational database systems in various application domains, developing top-k retrieval methods that work within the operational constraints is crucial. The query-mapping approach not only has operational advantages, but also provides effective mechanisms for range query estimation (Ayanso, Goes, and Mehta, 2007; Bruno et al., 2002; Chaudhuri and Gravano, 1999). Nevertheless, prior techniques for query-mapping are either simple heuristics that lead to significant performance variation by database setting (for example, Chaudhuri and Gravano, 1999) or workload-adaptive that requires learning of the data using a set of queries with pre-specified data characteristics (for example, Bruno et al., 2002; Chen and Ling, 2002).

This paper presents a cost-based query-mapping methodology that allows evaluation of cost tradeoff at the level of an individual query. The query-level method is an improvement over the workload-adaptive method which may lead to inefficient performance for queries that deviate in characteristics or distribution from pre-specified workloads. Our cost-based range estimation methodology accounts for cost and performance factors for each query and shows relatively consistent performance under different experimental settings.

The remainder of this paper is organized as follows. The following section reviews the literature on top-k querying with an emphasis on the query-mapping approach. The subsequent section presents the cost-based range estimation procedure and its model components and assumptions. This will be followed by the discussion of the experimental setting used to assess the performance efficiency of the cost-based strategy and the results obtained. The final sections discuss limitations and future research directions, and provide concluding remarks.

RELATED LITERATURE

The Web has proved to be an ideal example for the performance impact of top-k querying. As a result, document and multi-media retrieval has been the focus of most research in the IR field (Baeza-Yates and Ribeiro-Neto, 1999; Bertino, Rabbiti, and Gibbs, 1988; Bustos, Keim, Sauge, Schreek, and Vrančić, 2005; Chan, 2006; Chaudhuri, Gravano, and Marian, 2004; Fagin, 1998; Kwok and Zhao, 2006; Salton and McGill, 1983). Although RDBMSs primarily support exact matches to
Related Content

Public Sector Data Management in a Developing Economy
[www.igi-global.com/chapter/public-sector-data-management-developing/6208?camid=4v1a](www.igi-global.com/chapter/public-sector-data-management-developing/6208?camid=4v1a)

The Expert's Opinion
[www.igi-global.com/article/expert-opinion/51077?camid=4v1a](www.igi-global.com/article/expert-opinion/51077?camid=4v1a)

Semantic Integrity Constraint Checking for Multiple XML Databases
[www.igi-global.com/chapter/semantic-integrity-constraint-checking-multiple/7931?camid=4v1a](www.igi-global.com/chapter/semantic-integrity-constraint-checking-multiple/7931?camid=4v1a)

A Case Study of the Military Utility of Telemedicine
[www.igi-global.com/chapter/case-study-military-utility-telemedicine/8284?camid=4v1a](www.igi-global.com/chapter/case-study-military-utility-telemedicine/8284?camid=4v1a)