Decision Support Systems 57 (2014) 258–273

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Range query estimation with data skewness for top-k retrieval

Anteneh Ayanso [a,*], Paulo B. Goes [b], Kumar Mehta [c]

[a] Department of Finance, Operations, and Information Systems, Goodman School of Business, Brock University, 500 Glenridge Avenue, St. Catharines, ON L2S 3A1, Canada
[b] Department of Management Information Systems, Eller College of Management, University of Arizona, 1130 E. Helen Street, Tucson, AZ 85721, USA
[c] Department of Decision Science and MIS, School of Management, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

ARTICLE INFO

Article history:
Received 10 March 2013
Received in revised form 2 August 2013
Accepted 16 September 2013
Available online 23 September 2013

Keywords:
Top-k query
Query-mapping
Query processing
Cost model
RDBMSs

ABSTRACT

Top-k querying can significantly improve the performance of web-based business intelligence applications such as price comparison and product recommendation systems. Top-k retrieval involves finding a limited number of records in a relational database that are most similar to user-specified attribute-value pairs. This paper extends the cost-based query-mapping method for top-k retrieval by incorporating data skewness in range estimation. Experiments on real world and synthetic multi-attribute data sets show that incorporating data skewness provides a robust performance across different types of data sets, query sets, distance functions, and histograms.

1. Introduction

The integration of database and information retrieval technologies is becoming increasingly important, with more and more textual data being stored in relational database management systems (RDBMSs). Many customer-centric applications today require top-k retrieval in which the *k* most important objects are returned among the potentially huge answer space by a given ranking (distance) function [17]. RDBMSs primarily support queries that deal only with data that exactly match selection criteria [23]. As a result, end-users face the challenge of routinely specifying value ranges of attributes in search of a limited number of approximate matches.

Much of the existing research in top-k retrieval has focused on techniques that involve high-performance indexing that requires significant changes in the query engines of RDBMSs. Due to the complexity in incorporating new index structures in RDBMSs, developing scalable as well as practical top-k retrieval methods is crucial. Among the various approaches proposed in the literature, the query-mapping approach (see Refs. [1,2,4,5,9]) has shown that database profiles can be effectively used to estimate approximate range queries in order to avoid the requirement of a full sequential scan of a relation to answer top-k queries. The key advantage of this approach is that it can be operationalized at an application layer outside of core query engines of RDBMSs. As a result, the development of new data access methods and specialized index structures can further enhance its feasibility in various application environments [5]. Therefore, the practical appeal of the query-mapping approach provides significant opportunity if the underlying technical limitations are overcome through improvements on the range estimation procedure. This paper extends the cost-based query-mapping strategy [1,2] by incorporating data skewness in the range estimation procedure.

For the purpose of summarizing the contents of relations and estimating query result sizes for efficient execution plans, many current database systems (e.g. IBM DB2, Informix, Microsoft SQL Server, and Oracle) maintain some type of histograms. Histograms partition the underlying relations into distinct buckets by specifying value ranges of attributes and the number of database tuples (or records) available within those ranges. In histogram construction, a bucket is defined as a partition of a data set specified by attribute value ranges as boundaries and number of database tuples (or records) available within those ranges as frequencies. These summary statistics are key components of the query mapping approach for top-k retrieval. Another key component of the query-mapping approach is a distance function, which is used for measuring the "closeness" of tuples to a query point. Distance or scoring functions based on *l*-norms (e.g., *Euclidean*, *max*, and *sum*) are commonly used in top-k research [1,2,5,9]. From top-k retrieval perspective, an important property of the *l*-norm distances is *monotonicity* which states that if a tuple is closer along each attribute to the values of a query than some other tuple is, then, the distance from the query point to this tuple cannot be longer than that of any other tuple [5,9]. This allows measuring relevance and ranking of tuples to user queries. Thus, the query-mapping mechanism utilizes distance functions and histograms maintained in RDBMSs to obtain upper and lower bounds for selectivity estimates and determine an equivalent range query. Building upon this query-mapping strategy, the cost-based methodology [1,2] further leverages the trade-offs in query processing costs to determine

* Corresponding author. Tel.: +1 905 688 5550x3498.
E-mail addresses: aayanso@brocku.ca (A. Ayanso), pgoes@eller.arizona.edu (P.B. Goes), kmehta1@gmu.edu (K. Mehta).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.09.005

an optimal range for answering top-k queries. More specifically, the method incorporates the tradeoff between the cost of dealing with excess tuples and the cost of re-executing (restarting) a query when an estimated range fails to meet the desired number of tuples. Therefore, the method determines a cost-optimal range by minimizing the sum of the expected costs of re-processing and the expected costs of handling results in excess of the required number.

Unlike previous histogram-based methods, such as the dynamic workload-based strategy (DWBS) [5], the cost-based query-mapping method does not require training different workloads of queries for range estimation. However, like previous histogram-based methods, the method utilizes histogram information for range query estimation. The use of histograms as the basis for range query estimation is generally constrained by data distribution, where histograms are assumed to have uniform tuple density within individual histogram buckets or data partitions [4]. In the presence of data skewness, range estimation can be affected by the degree to which histogram buckets conform to uniform tuple density, particularly in multidimensional environments. When using the Uniform distribution assumption for histogram buckets, the result size estimation is based on the notion that tuples are evenly spread over a given search bound. However, in the presence of data skewness, the Uniform distribution assumption may affect the range estimation accuracy for top-k retrieval. This paper particularly addresses this common limitation of the histogram-based methods [1,2,5,9] by incorporating data skewness in histograms in the range estimation process and extending the modeling framework of the cost-based strategy [1,2].

Prior research in query optimization has shown that the use of the attribute value independence assumption for joint data distributions in multidimensional environments can lead to inaccurate result size estimations (e.g., see [11]). Consequently, there has been a significant amount of work on efficient techniques to approximate joint data distributions using multi-dimensional histograms (e.g., see Refs. [24–27]). The underlying assumption in these histogram techniques is that the distribution of values of a relation is described by an *n*-dimensional histogram. Because the objective of the query-mapping strategy is to estimate an equivalent range query for top-k retrieval, the implication of the attribute value independence assumption on approximating joint data distribution is equally significant.

Furthermore, in top-k query-mapping, the consideration of multiple attributes changes how the search space should be viewed in multi-dimensional histograms for result size estimation. For instance, in a range defined by a distance from a given query point, there can be several histogram buckets whose area/volume is either fully or partially included within this distance range. This affects the approach that can be taken in defining the search bound and determining the number of tuples available at and around a given query point. In the literature, there are two different approaches to define the search bound for a given query point. The first approach [5,9] considers buckets as atomic and takes an ordering of "optimistic" and "pessimistic" distances from the query point to the edges of each bucket. For multi-dimensional histogram, if buckets are not considered atomic, measuring the closeness of tuples in the different partitions of the data space requires estimating the volume overlap of buckets. This is mainly because there can be several buckets of tuples at and around a given query point in the multidimensional space. Thus, the second approach [4] involves the estimation of the volume overlap of buckets that fall within a given distance from a query point as well as the estimation of a data skewness parameter to adjust for the number of tuples that are expected to be included within this range. In measuring data skewness, prior research suggests the use of an *error* metric [13] or *deflation* parameter [4] that can be stored for each histogram or for each histogram bucket at the time of histogram construction, respectively. These parameters are then used in adjusting the estimation of the number of tuples available within a given range. It is important to note that the exact way of measuring volume overlap as well as data skewness in a multi-dimensional space are separate research issues. Given any available estimation techniques for volume overlap and data skewness, the objective of this paper is to develop a cost-based range estimation model that accounts for the deviation of data from the Uniform distribution assumption in a multidimensional histogram environment.

The rest of the paper is organized as follows. Section 2 provides a review of related research in top-k querying. Section 3 presents the cost-based estimation model, followed by the extension of the model with data skewness in Section 4. Section 5 describes the experimental setting, followed by the discussion of the computational results in Section 6. Finally, Section 7 provides concluding remarks and discusses limitations and future research directions.

2. Related literature

The Web has proved to be an ideal example for top-k querying. As such, document retrieval has been the focus of most research in the information retrieval (IR) field [3,6,16]. Top-k querying in multimedia systems is another area that stimulated related research in RDBMSs [14,21]. In the absence of efficient methods for exploratory search and retrieval in RDBMSs, the naïve approach requires an exhaustive scan of the database. This is obviously not a practical approach for many RDBMS applications which require more efficient and scalable performance. As a result, different streams of top-k processing techniques have been proposed in the recent literature [20].

The techniques in the existing literature are different, depending on whether they can be incorporated at the core of query engines or outside query engines at an application layer. Techniques that work at the core query engine level implement specialized rank-aware query operators or introduce new query algebra for query optimization [8,19,22]. On the other hand, the techniques that work at an application level use specialized indexes [7,28] or materialized views [12,18] to improve query response time during execution.

The techniques that are closely related to the method presented here belong to the query-mapping stream [1,2,4,5,9,10]. These techniques work at an application level and provide a mechanism to convert a top-k query into a conventional range query that RDBMSs support. In particular, these techniques avoid the need to do a full sequential scan of the database to answer top-k queries. Instead, these techniques estimate approximate range queries relying on summary information about the database in the form of histograms maintained in RDBMSs [1,2,4,5,9] or sample tuples obtained during query execution [10]. This paper belongs to this stream and extends the cost-based query-mapping methodology [1,2] by incorporating data skewness in range estimation. Table 1 summarizes the key contributions and limitations of the techniques that are relevant to this paper.

2.1. Estimating volume overlap

Before presenting the details of the range estimation model, we provide a review of the volume overlap and data skewness parameter estimation techniques in the literature. As discussed earlier, the query-mapping techniques utilize summary statistics about the database in the form of histograms, which create a partition of the data into distinct buckets of uniform tuple distribution to facilitate result size estimation. Thus, range estimation for top-k retrieval utilizes this histogram information and a distance function (e.g., *Euclidean*, *max*, and *sum*) to measure the "closeness" of database records to query conditions. For a range defined by a distance from a given query point, there can be several histogram buckets whose area/volume is either fully or partially included within this distance range. Consequently, if buckets are not considered atomic, measuring the closeness of tuples in the different partitions of the data space requires computing the volume overlap of buckets. The method we adopt for computing the volume overlap of buckets has a closed form solution for the *max* distance function and an approximate solution for the *Euclidean* and *sum* distance functions