Computers and Mathematics with Applications 57 (2009) 1037–1047

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

A new approach to building histogram for selectivity estimation in query processing optimization[☆]

Xin Lu[a], Jihong Guan[b,*]

[a] Department of Computer Sci. and Eng., Fudan University, Shanghai 200433, China

[b] Department of Computer Sci. and Techl., Tongji University, Shanghai 200092, China

ARTICLE INFO

Keywords:
Selectivity estimation
Histogram
Query optimization

ABSTRACT

Recently, histograms have been considered as an effective way to produce quick approximate answers to decision support queries. They are also taken as a basic tool for data visualization and analysis. In this paper, we propose a new approach to constructing histograms for selectivity estimation in query processing optimization. Our approach uses a new criterion, i.e., aggregate error minimization, to direct the construction of the target histogram. We develop the algorithm of aggregate error minimization based histogram construction, and demonstrate the effectiveness and efficiency of the proposed approach by experiments over both real-world and synthetic datasets.

1. Introduction

In database systems, the query processing optimizer is an essential module that generates the plan of how a query will be executed. Selectivity estimation plays a key role in the query optimization process. Inaccurate selectivity estimation may lead to the choice of a suboptimal execution plan, and the execution times of an optimal plan and a suboptimal plan may be substantially different.

Consider a relation R, an attribute A of R, and the domain D of A. Suppose there are n tuples in R, and these tuples' values of attribute A are $x_1, x_2, \ldots, x_n$. For an arbitrary value a or interval r in D, the *selectivity estimation* task is to evaluate how many tuples whose values of attribute A are equal to a or fall in r. To this end, we should have the exact or an approximate distribution of attribute A's values over domain D.

There exist a number of statistical methods used in commercial databases for the purpose of selectivity estimation. In the database academia, a lot of effort has also been put on this problem over the past two decades, and a number of approaches have been proposed, including histogram-based approaches [1–14], sampling-based approaches [15,16], and wavelet-based synopses [17] etc.

Histogram is one of the most natural and useful forms of data representation for the purpose of data summary and analysis; it is also a very popular and flexible way to track data distribution in databases. As a research topic that has been studied extensively, quite a number of algorithms have been proposed for efficiently constructing histograms over a single attribute [3–8] or multiple attributes [9–14]. Given a dataset and a fixed amount of storage space, a histogram is constructed to optimize a certain goal function. Usually, it is used to minimize the selectivity estimation error caused by the discrepancy between the built histogram and the real data distribution. In other words, the goal is to make the histogram be as close as possible to the real data distribution.

[☆] This work is supported by the National Natural Science Foundation of China under grants — no. 60496327, no. 60573183 and no. 60773123; Jihong Guan is also supported by the Program for New Century Excellent Talents in University of China (NCET-06-0376).

[*] Corresponding address: Department of Computer Science & Technology, School of Electronics and Information, Tongji University, Shanghai 200092, China. Tel.: +86 21 65109899; fax: +86 21 55664298.

E-mail addresses: luxin@fudan.edu.cn (X. Lu), jhguan@mail.tongji.edu.cn (J. Guan).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.10.056

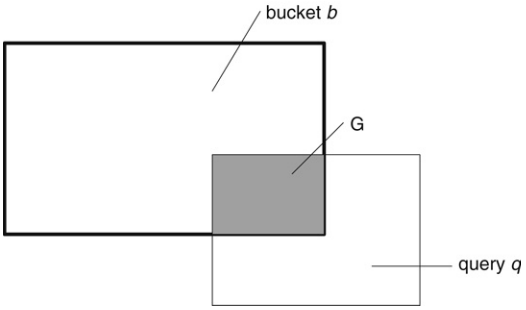

Fig. 1. Illustration of estimating the selectivity inside a bucket.

In this paper, we propose a new concept of error, *aggregate error*, for histogram construction. The *relative aggregate error* is used in histogram construction. Algorithm is developed to build histograms by minimizing the error value. Experiments over both real-world and synthetic datasets are carried out to evaluate the performance of the proposed method.

The remaining of this paper is organized as follows. Section 2 surveys related work of selectivity estimation. Section 3 states the problem of histogram construction. Section 4 introduces the approach to building histograms based on aggregate error minimization. Section 5 presents performance evaluation. And finally Section 6 conclude the paper and highlights future work.

2. Related work

In the database community, selectivity estimation is an important research issue of query processing optimization. Up to now, a number of techniques have been proposed for selectivity estimation, including histogram-based approaches, sampling-based approaches, and wavelet-based technique etc. Considering that this paper is about histogram construction, we survey mainly the approaches of histogram-based selectivity estimation.

Wavelet-based technique is a parametric approach, which was proposed as the replacement to histograms [17]. It uses wavelet decomposition for approximation. Wavelet decomposition is, roughly, the decomposition of a function into wavelets with hierarchical levels of detail. Wavelets are functions that look like decaying oscillating waves. There exist many kinds of wavelets, the most popular are Haar and linear ones. One wavelet represents the rough picture of data distribution over the entire domain. Wavelet-based approaches are inefficient because of the large amount of computation of matrix transformation.

Sampling-based methods have some advantages as follows: (1) They need not collect, store and maintain any statistical information; (2) They do not require imposing any a priori assumption on data distribution; (3) They are suitable for high-dimensional data. For these reasons, sampling-based methods for selectivity estimation have attracted considerable attention of database researchers. An exhaustive overview of classic sampling methods was presented as a survey by Olken and Rotem [15], which covers the research conducted in this area up to the year of 1990. And a more in-depth study on random sampling for use in database systems was given in Olken's Ph.D. dissertation [16].

Histograms are the most commonly used form of statistics in practice because they incur almost no run-time overhead and are effective even with a very small amount of storage space. They have been used in DB2, Oracle, and Microsoft SQL Server. Several types of histograms have been proposed and studied, including MinSkew [18], EquiHeight [21], GENHIST [20], STGrid [22], STHoles [9].

Minskew is a histogram construction algorithm originally proposed for selectivity estimation of range queries on non-uniform datasets [18]. At first, Minskew treats the whole data space as a bucket, then it partitions the bucket into a set of small buckets, each of which does not intersect with any other bucket, and the union of all buckets covers the entire data space. Each bucket b_i contains the number $b_i.num$ of objects that fall inside bucket b_i. Fig. 1 shows a query q that intersects with bucket b in two-dimensional space. The gray area corresponds to the intersection between bucket b and the query q. The expected number of objects in the intersection region G of b and q is estimated as $b.num * area(G)/area(b)$, where $area(G)$ and $area(b)$ are the areas of the intersection region G and bucket b, respectively. Generally, the estimated selectivity is obtained by summing the results of all intersecting buckets.

Since the estimation algorithm above applies uniform assumption within a bucket, the more uniform the intra-bucket distribution is, the more accurate the estimation will be. Minskew defines the spatial-skew (denoted as b.skew) for a bucket b as the variance of the spatial densities of all points inside it. It is obvious that a partitioning with small spatial-skewness is likely to be highly accurate in approximating the given data. So Minskew aims at minimizing the weighted sum of spatial-skews of all buckets. Unfortunately, building the optimal partitioning is NP-Hard [19]. To reduce the complexity of constructing good partitionings, Minskew partitions the original space to a grid with $H * H$ regular cells (where H is a parameter), and records the number of objects that fall in cell c as $c.num$. Thus, $b.num$ is equal to the sum of objects falling in all cells covered by bucket b. It then uses a greedy approach to seek a local optimal result for reducing computation cost.

EquiHeight [21] is an approach to constructing equi-height histograms. In an equi-height histogram, each bucket contains approximately the same number of data objects. Considering that data is usually unevenly distributed, a dense area will be