Information Sciences 188 (2012) 121–150

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

A probabilistic framework for estimating the accuracy of aggregate range queries evaluated over histograms ☆

Francesco Buccafurri [a,*], Filippo Furfaro [b], Domenico Saccà [b]

[a] DIMET, University Mediterranea, 89100 Reggio Calabria, Italy
[b] DEIS, University of Calabria, 87036 Rende, Italy

ARTICLE INFO

Article history:
Received 17 February 2009
Received in revised form 29 April 2011
Accepted 9 November 2011
Available online 18 November 2011

Keywords:
Multidimensional data
Compression
Histogram
Range query

ABSTRACT

A histogram over a multi-dimensional data set is a synopsis consisting of aggregate data summarizing the values of the points inside non-overlapping ranges of the domain. Owing to their effectiveness in supporting a fast (though approximate) estimation of the answers of aggregate range queries, histograms are widely used in several contexts dealing with multi-dimensional data, especially those where the precision of the answers (within reasonable limits) is not the major requirement. However, the practical impact of histograms has been limited by the fact that, so far, no mechanism has been defined which provides a reliable (non-trivial) quantification of the degree of approximation of the query estimates. In this paper, this problem is addressed by introducing a probabilistic framework which allows for estimating the accuracy of the approximate answers resulting from evaluating aggregate queries over a histogram. Specifically, given a histogram over a data set, the answer of an aggregate range query is modeled as a random variable, whose probability distribution depends on the type and the values of the aggregate data stored in the histogram. Therein, the mean value and the variance of this random variable represent an estimate of the actual answer of the corresponding query and of the error rate, respectively. The proposed framework can exploit different kinds of aggregates (namely, *sum* and *count*) stored in the histogram, as well as integrity constraints defined over the original data.

1. Introduction

In many application contexts, such as statistical databases, transaction recording systems, scientific databases, query optimizers, OLAP (On-line Analytical Processing), and many others, a multidimensional view of data is often adopted: the attributes of the tuples are partitioned into two disjoint sets, namely *dimensional* and *measure* attributes, and data are logically represented as points in a multidimensional space. Specifically, each point is associated with a set of coordinates and a set of measures, whose values correspond to the values of the dimensional attributes and the measure attributes, respectively. This representation model supports data exploration and aggregation, as it makes the specification of the range of the data domain of interest an easy and intuitive task. Both data exploration and aggregation are performed by posing aggregate *range queries*, i.e., queries asking for aggregate information on the data inside a range, such as the number of points or the sum of their values.

☆ An abridged version of this paper appeared in [4].
* Corresponding author.
E-mail addresses: bucca@unirc.it (F. Buccafurri), furfaro@si.deis.unical.it (F. Furfaro), sacca@unical.it (D. Saccà).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.11.009

In demanding applications, multidimensional data are often summarized into lossy synopses of aggregate values (called *histograms*), and range queries are executed over these aggregate data, without accessing the original ones. Basically, a histogram is built by partitioning the data domain into disjoint regions (called *buckets*) and then storing, for each bucket, some aggregate information (such as the sum of the values inside its range) instead of the detailed distribution of the data covered by the bucket itself. Range queries are then evaluated by performing suitable interpolations on the aggregate data associated with the buckets overlapping the range of the query. Owing to the smaller amount of information represented in the histogram compared with the raw data, it holds that:

- on the one hand, accessing the histogram rather than the original data enables a fast estimation of answers, thanks to the reduced amount of data which must be accessed to estimate the answer;
- on the other hand, since applying summarization yields a loss of information, query answers recovered from the histogram are approximate.

Hence, histograms are employed to support query answering in the contexts where the exactness of query answers is not a requirement, and the preferable choice is that of avoiding long waits which would yield a precision which is often unnecessary. Typical examples of these contexts are *query optimization* and *data exploration in Decision Support Systems* (DSSs). As regards the former, query optimizers in RDBMSs can build an effective query evaluation plan by estimating the selectivity of intermediate query results [9], which can be accomplished by retrieving aggregate information on the frequencies of attribute values. Thus, in this case, histograms are built on the frequency distribution of the attribute values occurring in database relations. Obviously, the execution plan for a given query is effective only if it can be computed efficiently, that is if building the plan takes much less time than answering the query itself: thus fast computation of aggregations is mandatory. Moreover, a dramatic precision in evaluating aggregates is not needed, as knowing the order of magnitude of the selectivity of intermediate queries suffices to build an effective execution plan. As regards the DSS context, users often perform preliminary explorations of the data (organized according to a multidimensional structure called *datacube* [8,22,27]), in order to find the portions where a more detailed analysis is needed. In this scenario, high accuracy in less relevant digits of query answers is not needed, as providing their order of magnitude suffices to locate the regions of the database containing relevant information. At the same time, fast answers to these preliminary queries allow users to focus their explorations quickly and effectively, thus saving large amounts of system resources.

In the last few years, many research efforts have been devoted to refining histogram construction techniques and improving their "effectiveness", in terms of the capability of histograms of providing query estimates with reasonable error rates (the reader is referred to *related work* section for a summary of these approaches). Indeed, so far, no mechanism has been defined for providing a reliable (non-trivial) quantification of the degree of approximation of the query estimates retrieved from histograms (except from [30], where only a rough evaluation strategy for upper bounds of error rates was introduced). This gap has seriously limited the practical impact of histograms, precluding their use in several contexts where approximate query answers are useful. For instance, in the DSS scenario, the possibility of getting a reasonably tight estimation of the error rates could make histograms usable not only for preliminary explorations, but also for efficiently obtaining "definitive" analysis reports. As an example, consider an analyst who posed a query on a histogram (summarizing the financial data of her enterprise) and receives an approximate answer along with a reliable estimate of the error rate. In this case, if the analyst estimates that the error rate is "tolerable" (in the sense that it gives a picture of the real world which is accurate enough for her analysis purposes), the analyst may decide to write reports and/or suggesting decisions relying on the approximate answers only, with no need of the exact answers.

1.1. Our contribution

This work aims at plugging this gap of the research literature: we address the problem of providing reliable estimates of the error rates of the approximate query answers retrieved from histograms. In this regard, we define a probabilistic framework which enables the error rates of the approximate answers to be estimated for the two types of aggregate queries (namely, *count* and *sum* queries) which are generally posed against histograms. Our framework is likely to have a strong impact on the practical use of histograms, as it makes no assumption on how the histogram has been obtained, and it works on every histogram built using any construction technique in the literature.[1]

In more detail, our framework relies on modeling query answers as random variables, according to the following rationale. Given a multidimensional data set M and a histogram S summarizing M, we denote the transformation from M to S by τ, thus $S = \tau(M)$. Let $\mathcal{M}_S$ denote the set of all the data sets $\overline{M}$ such that $\tau(\overline{M}) = S$, thus any data set $\overline{M} \in \mathcal{M}_S$ is a possible guess of M, on the basis of the knowledge of S. Thus, if we are given a range query Q on M, estimating Q from S can be viewed as guessing the answer of Q on M by applying the range query Q to any $\overline{M}$ in $\mathcal{M}_S$. According to this observation, we will model the estimate of the range query Q posed against S as a random variable ranging over the set of values consisting of the answers resulting from evaluating Q over every $\overline{M}$ in $\mathcal{M}_S$. In order to analyze the estimation error, we will study this random

[1] We refer to histograms where bucket boundaries define a partition of the data domain in a "strict" sense (that is, bucket overlapping is not allowed). Although the term "histogram" is also used to denote synopses where buckets may overlap, our assumption is not a severe restriction, since many histograms are based this kind of partition.