Data & Knowledge Engineering 88 (2013) 206–223

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Bichromatic buckets: An effective technique to improve the accuracy of histograms for geographic data points

Hai Thanh Mai *, Jaeho Kim, Myoung Ho Kim

Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Available online 24 April 2013

Keywords:
Database theory
Spatial data processing
Query optimization
Selectivity estimation
Histograms

ABSTRACT

Histograms have been widely used for estimating selectivity in query optimization. In this paper, we propose a new technique to improve the accuracy of histograms for two-dimensional geographic data points that are used in many real-world Geographic Information Systems. Typically, a histogram consists of a collection of rectangular regions, called buckets. The main idea of our technique is to use a straight line to convert each rectangular bucket to a new one with two separating regions. The converted buckets, called *bichromatic* buckets, can approximate the distribution of data objects better while preserving the simplicity of originally rectangular ones. To construct bichromatic buckets, we propose an adaptive algorithm to find good separating lines. Two strategies to find the separating lines, one based on the potential skewness gains of the candidate lines and the other based on the difference of density levels of the data regions, are proposed and used flexibly within our algorithm. Then, we describe how to apply the proposed technique to existing histogram construction methods to improve the accuracy of the constructed histograms further. Results from extensive experiments using real-life data sets demonstrate that our technique improves the accuracy of the histograms by 2 times on average.

1. Introduction

In databases, estimating the selectivity of queries is an essential part of query optimization [1–3]. Accurate selectivity estimates can help the query execution engine to choose the most efficient query plan. Therefore, over the last decades, the problem of selectivity estimation has been intensively investigated. Several selectivity estimation approaches have been proposed, such as histograms [4–14], wavelet transformation [15,16], singular value decomposition [5], discrete cosine transform [17], kernel estimators [18,11], and sampling [19–21]. Among these approaches, histograms have been shown to be one of the most popular and effective ways to obtain accurate estimates of selectivity [22,11,13].

Let D be a data set of our interest and S be the data space of D. A histogram H for D consists of a set of m buckets B_i $(1 \leq i \leq m)$ where m is usually a system parameter. Each bucket B_i has a data space S_i that is a subspace of S and a frequency F_i that is the number of data objects in S_i. The data space S_i is an interval, a rectangle, or a hyper-rectangle if the data objects have one, two, or higher than two dimensions, respectively. With these m buckets, H approximates the distribution of the data in D. Now, suppose that a query Q on D is given by the user to retrieve data objects within a range S_Q, an estimate of the selectivity of Q (i.e., the number of data objects in S_Q) by using the histogram H, denoted by $F_Q(H)$, is typically computed as

$$F_Q(H) = \sum_{i=1}^{m} \frac{|S_i \cap S_Q|}{|S_i|} \cdot F_i \tag{1}$$

* Corresponding author. Tel.: +82 42 350 7730; fax: +82 42 350 2255.
E-mail addresses: mhthanh@dbserver.kaist.ac.kr (H.T. Mai), jaeho@dbserver.kaist.ac.kr (J. Kim), mhkim@dbserver.kaist.ac.kr (M.H. Kim).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.04.003

under the *intra-bucket uniform distribution assumption*. Here, || denotes the size of data space and $(S_i \cap S_Q)$ denotes the intersecting area of S_i and S_Q. Note, however, that the details of $F_Q(H)$ may differ, depending on the histogram methods. Although the uniform distribution of data inside the buckets is important for accurate selectivity estimation, it is well-known that such organization of the buckets is computationally intractable [23].

In this work, we study the problem of constructing highly accurate histograms for selectivity estimation in spatial databases. We focus on the histograms for two-dimensional geographic data points where updates do not frequently occur. Data in this form are widely used in Geographic Information Systems (GISs). The histogram must be constructed so that its estimated selectivity for the query must be close to the true selectivity of the query as much as possible. However, creating an accurate histogram for multi-dimensional data, including two-dimensional geographic data points, is not an easy task. When the region of a query fully covers the region of a bucket *B*, we can use *B*'s object frequency directly. In contrast, when the region of a query partially overlaps with or is fully contained in the region of *B*, the problem may arise. In these latter cases, the estimated selectivity value for the overlapping region between the query and *B* is computed in proportion to the size of this overlapping region. Here, if data objects are distributed uniformly within *B*, our estimation is close to the real object frequency. Otherwise, we are very likely to obtain wrong results. For instance, let us consider a bucket *B* and a query *Q* shown in Fig. 1. The size of the overlapping area between *Q* and *B* (i.e., the gray area in the figure) is 1/4 of the size of *B*. If uniform distribution of objects is assumed, the estimated selectivity of this overlapping region is 1/4 of the object frequency of *B*, i.e., 10. Nevertheless, since objects in *B* are not uniformly distributed and most of them lie at the lower-left part of the bucket whose region does not overlap with *Q*, the estimate 10 is far from the correct number 1.

In real-life data sets, as the uniformity is rare and non-uniformity is naturally popular, many histogram construction methods have tried to address the skewness (i.e., non-uniform distribution) problem of the data, e.g., MinSkew [7], GenHist [11], RkHist [13], and STHist [14]. These methods differ from each other in the ways they allocate rectangular buckets onto the data space, so that the data distribution in each bucket is close to uniformity as much as possible. Nevertheless, we have observed that there are many regions, such as the region illustrated in Fig. 1, where it is very difficult to improve the uniformity of data distribution in the bucket further. The reason is that the bucket is a rectangle while the data distribution may have many different shapes. One straightforward solution is to allocate many more buckets to such complex regions. Nevertheless, allocating more buckets to one region means that fewer buckets can be used for other regions because the bucket quota is limited. Another solution is to use generally polygonal shapes for the buckets instead of rectangles. Polygons can fit the distribution of the data objects better. However, since a much higher amount of memory must be used to describe the polygons than the rectangles in general, much fewer buckets can be used. Moreover, this solution incurs higher complexity than the traditional rectangle-based solution in partitioning the data space and deciding the specific shapes of the polygons.

In this paper, we propose a new technique to improve the accuracy of the histograms. Notice that, this technique is not a stand-alone histogram construction method. Instead, it serves as a useful component to be added into existing histogram construction methods to enhance the accuracy of the histograms constructed by these methods. Our main idea is to use a straight line to convert each rectangular bucket that is constructed by an existing histogram method to a new one with two separating regions. The converted buckets, called *bichromatic* buckets, can approximate the objects' distribution better while preserving the simplicity of the originally rectangular ones.

For converting original buckets to bichromatic ones, we propose an algorithm to find good separating lines. Two strategies are used adaptively within this algorithm to ensure that the algorithm is both effective and efficient. When the distribution of data objects inside a bucket is too complicated, several possible choices of the separating lines are examined and the best one is chosen based on a notion of potential skewness gain. On the contrary, when the object distribution inside the bucket is moderately linearly separable (i.e., not much complicated), statistical tools are used to find the separating line directly, based on the difference of density levels among the sub-regions of the bucket.

As we mentioned above, the proposed technique takes the role of an additional component into existing histogram methods. Hence, we present how to apply this technique to existing histogram methods. The integration of our technique to existing

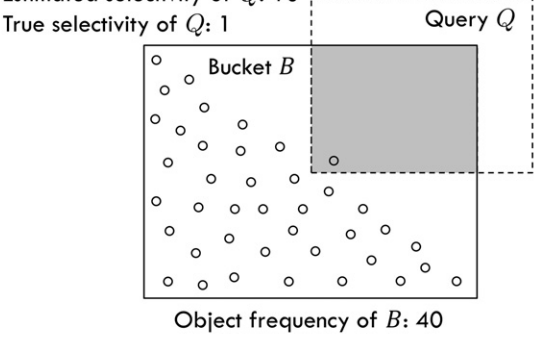
Estimated selectivity of Q: 10
True selectivity of Q: 1
Query Q
Bucket B
Object frequency of B: 40

Fig. 1. Inaccuracy of the histogram when data points in the bucket are not uniformly distributed.