Spatial Selectivity Estimation for Intersection region Information Using Cumulative Density Histogram

byung Cheol Kim
Database Laboratory, Chungbuk National University,
12 gaeshin-dong heongduk-gu, Cheongju, Chungbuk, 361-763, korea
bckim@ok.ac.kr

Kyung Do Moon
Database Laboratory, Chungbuk National University,
12 gaeshin-dong heongduk-gu, Cheongju, Chungbuk, 361-763, korea
kdmoon@chol.com

Keun Ho Ryu
Database Laboratory, Chungbuk National University,
12 gaeshin-dong heongduk-gu, Cheongju, Chungbuk, 361-763, korea
khryu@dblab.chungbuk.ac.kr

ABSTRACT

Multiple-count problem is occurred when rectangle objects span across several buckets. The Cumulative Density (CD) histogram is a technique which solves multiple-count problem by keeping four sub-histograms corresponding to the four points of rectangle. Although it provides exact results with constant response time, there is still a considerable issue. Since it is based on a query window which aligns with a given grid, a number of errors may be occurred when it is applied to real applications. In this paper, we proposed selectivity estimation techniques using the generalized cumulative density histogram based on two probabilistic models: (1) probabilistic model which considers the query window area ratio, (2) probabilistic model which considers intersection area between a given grid and objects. In order to evaluate the proposed methods, we experimented with real dataset and experimental results showed that the proposed technique was superior to the existing selectivity estimation techniques. The proposed techniques can be used to accurately quantify the selectivity of the spatial range query on rectangle objects.

Keywords: Selectivity Estimation, Spatial Histogram

1. Introduction

Recently, with the advance of computing technology, the developments of GIS applications have received considerable attention. Analysis of the performance of query in a spatial database management system (SDBMS) becomes more essential due to the emergence of these GIS applications. To analyze performance of query efficiently, estimating the result size for spatial queries accurately is crucial [4].

Several techniques have been proposed in the literature to estimate query result sizes, including histograms, sampling and parametric techniques [1,4,5,6]. Of these, histograms approximate the frequency distribution of an attribute by grouping attribute values into buckets and approximating true attribute values and their frequencies in data based on summary statistics maintained in each bucket [2,3,4,5,6].

This paper focuses on estimating the selectivity of range queries on rectangular objects. Retangular objects incur multiple-count problem when they span across the several buckets. To solve this multiple-count problem, the CD(Cumulative Density) and Euler histograms are proposed in the literature [4,5]. Those techniqes give very accurate results on rectangular objects. To apply the histogram techniques to real applications, we should make no assumptions on data or queries. The CD hisgoram gives a good result on both point and rectanglular objects while the Euler histogram can be just applied to region objects. Although the CD histogram gives very accurate results on spatial datasets, it may not be universally applicable because they are based on an assumption that query window aligns with a given grid cell.

Motivated by the above reasoning, we generalize the CD histogram approach to handle query windows that do not align with a given grid. Our technique is based on two probabilistic models. Our experiments show that the generalized CD histogram with no assumptions can be applied to real applications.

2. Related Work

Selectivity estimation is a well-studied problem for traditional data types such as integer. Histograms are most widely used form for doing selectivity estimation in relational database systems. Many different histograms have been proposed in the literature and some have been deployed in commercial RDBMSs. However, selectivity estimation in spatial databases is a relatively new topic, and some techniques for range queries have been proposed in the literature [2,3,4,5].

In [3], Acharya et. al. proposed the MinSkew algorithm. The MinSkew algorithm starts with a density histogram of the dataset, which effectively transforms region objects to point data. The density histogram is further split into more buckets until the given bucket count is reached or the sum of the variance in each bucket cannot be reduced by additional splitting. In result, the MinSkew algorithm constructs a spatial histogram to minimize the spatial-skew of spatial objects. The CD Histogram is proposed in [5]. Typically when building a histogram for region objects, an object may be counted multiple times if it spans across several buckets. The CD algorithm addresses this problem by keeping four sub-histogram to store the number of corresponding corner points that fall in the buckets, so even if a rectangle spans several buckets, it is counted exactly one in a each sub-histogram. The Euler Histogram is proposed in [4]. Although many works have been performed in this area, most previous works have made certain simplifying assumptions about datasets and/or queries to keep the analysis tractable. As a result, they may not be universally applicable. Therefore, to develop more universally applicable selectivity estimation techniques, we generalize the CD histogram. The Generalized CD histogram has no any assumptions on datasets, and gives very accurate results on rectangular datasets.

3. Generalized Cummulative Density Histogram

In this section, we generalize the CD histogram approach to handle query windows that do not align with a given grid. We first descirbe the CD Histogram Framework, and then describe two probabilistic models that are used to generalize the CD histogram.

3.1 The CD Histogram Framework

Given the level of gridding h, the CD Histogram constructs a grid cell of 4h, and keeps the buckets corresponding to the grid cell.

For a query window Q(xa, ya, xb, yb), the number of objects S'(Q) that intersect the query Q can be calculated as follows:

$$S(Q)=\begin{pmatrix} H_{ll}(xb,yb)-H_{lr}(xa-1,yb)-H_{ul}(xb,yb-1) \\ +H_{ur}(xa-1,ya-1) \end{pmatrix}*P \quad (1)$$

where P is the probability which is used to generalize the CD histogram, and will be described in section 3.2 and 3.3.

Accuracy of the estimated selectivity is measured by the Average Relative Error AE, which is defined as follows:

$$AE=\frac{(\sum_{q_i\in Q}|S'(qi)-S(qi)|)}{\sum_{q_i\in Q}S(qi)}*100 \qquad (2)$$

where qi is a query of ith, S(qi) is the actual selectivity for the query qi. In the experiment, we used 20 query windows which are constructed ramdomly.

3.2 Generalization of CD Histogram using query window ratio

To develop more universally applicable selectivity estimation techniques, we should make little or no assumptions about the dataset and query window. To generalize the CD histogram, we first consider the probabilistic model based on query window area ratio which is used in MinSkew Histogram technique.

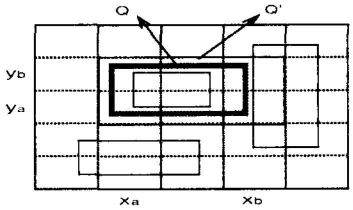


Fig. 1. Query Window Q and Q'

Fig. 1 illustrates query window Q and query window Q' extended to estimate the selectivity using CD histogram. The probabilistic model which considers the query window area ratio is defined as follows:

$$P=\frac{Area(Q)}{Area(B_Q)} \qquad (3)$$

With using probabilistic model P defined by equation 3, equation 1 can be written as follows:

$$S(Q)=\begin{pmatrix} H_{ll}(xb,yb)-H_{lr}(xa-1,yb)-H_{ul}(xb,yb-1) \\ +H_{ur}(xa-1,ya-1) \end{pmatrix}*\frac{Area(Q)}{Area(B_Q)} \quad (4)$$

3.3 Generalization of CD Histogram using intersection area

The common apporach for histogram is to keep a set of parameters which summarize the information about each dataset, and rely on a probabilistic model to provide an estimated result. In this section, we generalize the existing CD histogram using the probabilistic model which is based on the area of the intersection region between the objects and the grid cell. To estimate the selectivity using the probabilistic model, we keep additional information about the intersection area in each bucket. Therefore each bucket of the generalized CD histogram keeps five values H$_{ll}$, H$_{lr}$, H$_{ul}$, H$_{ur}$, and iArea, where iArea is the area of the intersection region between the objects and the grid cell.

An example of the generalized CD histogram keeping area of intersection regions is shown in Figure 2. The probabilistic model using the intersection area is defined as follows:

$$P=\frac{\sum_{i=k}^{i=l}\sum_{j=m}^{j=n}iArea(i,j)*Area_{i,j}(Q)}{\sum_{i=k}^{i=l}\sum_{j=m}^{j=n}iArea(i,j)} \qquad (5)$$

where k,l are the position values of x axis and n,m are the position values of y axis intersecting a query window. iArea(i,j) is the area of the intersection region between cell(i,j) and objects. Area$_{i,j}$(Q) is the area of the intersection region between cell(i,j) and query window.

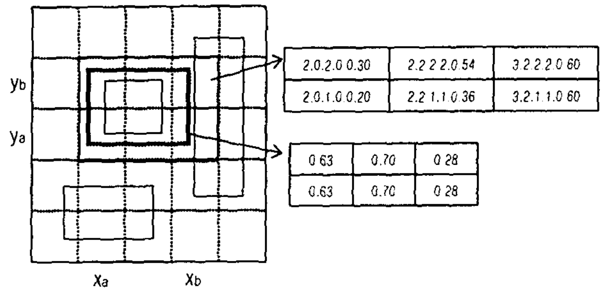


Fig. 2. an example of the Generalized CD Histogram

With using probabilistic model P defined by equation 5, equation 1 can be written as follows:

$$S'(Q)=\begin{pmatrix} H_{ll}(xb,yb)-H_{lr}(xa-1,yb)-H_{ul}(xb,yb-1) \\ +H_{ur}(xa-1,ya-1) \end{pmatrix}$$

$$*\,\frac{\sum_{i=k}^{i=l}\sum_{j=m}^{j=n} iArea\,(i,j)*Area_{i,j}(Q)}{\sum_{i=k}^{i=l}\sum_{j=m}^{j=n} iArea\,(i,j)} \qquad (6)$$

4. Experimental Results

To evaluate the effectiveness of the generalized CD histogram, we compare it with the MinSkew Histogram, Wavelet Histogram. In this section, we describe our experimental setup and report the performance results.

4.1 Datasets and Query

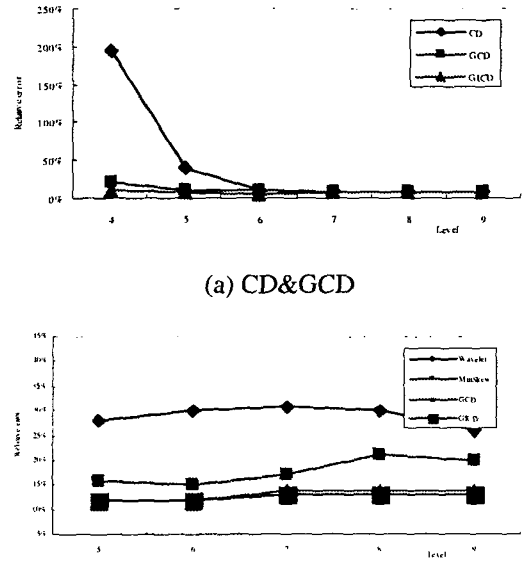
Our experiments were conducted on Intel Pentium IV 2GHz PC with dataset of commercial building located in Seoul Korea which contains 11,000 rectangles. To evaulate the proposed technique, we consider an estimation error, which is the difference between the selectivity that is estimated by the techniques and the actual selectivity.

4.2 Experimental Results

We describe the experiments that we performed to evaluate the performance of the proposed technique with respect to the grid granularity in this section.

The estimation accuracy as the grid level is shown in Fig. 3. The x and y-axis represent the grid level of histogram and average relative error respectively. We have obtained results for each technique using different levels (h=4 or 5,6,7,8,9).

Fig. 3(a) shows the results of the existing CD histogram and the proposed techniques, where CD is the conventional CD histogram, GCD is the generalized CD histogram based on probabilistic model using query window ratio, and GICD is the generalized CD histogram based on probabilistic model using intersection area.



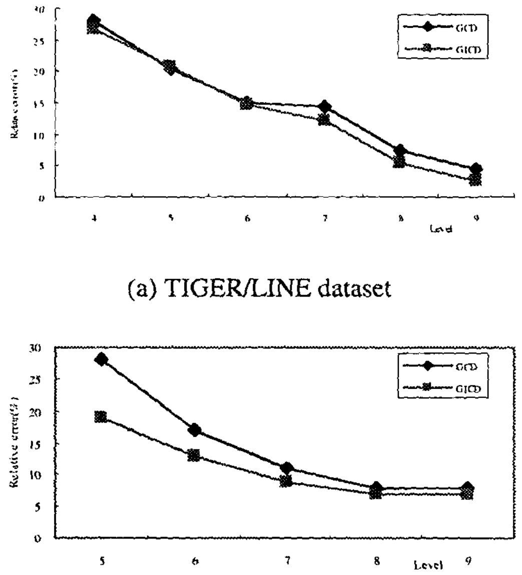
(a) CD&GCD

(b) GCD&Others

Fig. 3. Average Relative Error

This result shows that many estimated errors occur when the CD histogram is applied to handle query windows that do not align with a given grid. Fig. 3(b) shows the results of the existing techniques, Wavelet[3] and MinSkew[4], and the proposed techniques. The result shows that the generalized CD histogram gives much better estimates than the existing techniques which based on data transformation.

Fig. 4 (a) and (b) show the results of our technique on D2 and D3 dataset. The general trend of the graph is that the estimation accuracy improves as the gridding level increases. This is because probability which aligns query window to boundary of grid cell increases as the gridding level increases. We also find that GICD gives very good estimates over all the datasets as the previous results.



(a) TIGER/LINE dataset

(b) Sequoia Benchmark dataset

Fig. 4. Average Relative Error of the proposed method under varying data size

5. Conclusions

The selectivity on retanglular datasets may not be estimated accurately, because multiple-count problem occurs when rectangle objects span across several buckets. The CD histogram proposed to solve this problem gives very accurate results in a short time. However, a number of errors may occur when it is applied to real applications, since this approach is based on an assumption that a query window aligns with a given grid.

Therfore, in this paper we generalized the CD histogram approach using two probabilistic model, (1) probabilistic model which considers the query window area ratio, (2) probabilistic model which considers intersection area between a given grid and objects, to handle query windows that do not align with a given grid. With a real datasets, this paper has shown that the generalized CD histograms give a very accurate results with errors that are much lower than the previously histogram approaches.

In the future, we will improve the performance of the proposed method by exploring alternative probablistic models. We are also interseted in extending the CD histogram approach to more complicated queries such as spatial join with query window.

References

[1] Alberto Belussi, Christos Faloutsos, "Estimating the Selectivity of Spatial Queries using the `Correlation ' Fractal Dimension", InProc. 21st Int. Conf. Very Large Data Bases (VLDB), November, 1995, pp. 299-310

[2] Yossi Matias, Jeffrey Scott Vitter, Min Wang, "Wavelet-Based Histograms for Selectivity Estimation",In Proc. ACM SIGMOD Int. Conf. on Management of Data, 1998, pp.448-459

[3] Swarup Acharya, Viswanath Poosala, Sridhar Ramaswamy, "Selectivity estimation in spatial databases", In Proc. ACM SIGMOD Int. Conf. on Management of Data, 1999, pp.13-24

[4] C. Sun, D. Agrawal and A. El Abbadi, "Exploring Spatial Datasets with Histograms", , In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2002, pp.93-102

[5] Jin, N. An, A. Sivasubramaniam, "Analyzing Range Queries on Spatial Data", In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2000, pp. 525-534

[6] Ning An, Zhen-Yu Yang, Sivasubramaniam, A., "Selectivity estimation for spatial joins", In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2001, pp.368-375