Selectivity Estimation for Spatial Databases

Jeong Hee Chi, Jin Yul Lee, Keun Ho Ryu
Database Laboratory, Chungbuk National University
Cheongju Chungbuk, 361-763, Korea
{ jhchi, jinylee, khryu}@dblab.cbu.ac.kr

Abstract: Selectivity estimation for spatial query is curial in Spatial Database Management Systems(SDBMS). Many works have been performed to estimate accurate selectivity. Although they deal with some problems such as *false-count*, *multi-count* arising from properties of spatial dataset, they can not get such effects in little memory space Therefore, we need to compress spatial dataset into little memory. In this paper, we propose a new technique called *MW Histogram* which is able to compress summary data and get reasonable results. Our method is based on two techniques:(a)MinSkew partitioning algorithm which deal with skewed spatial datasets efficiently (b) Wavelet transformation which compression effect is proven. We evaluate our method via real datasets. The experimental result shows that the MW Histogram has the ability of providing estimates with low relative error and retaining the similar estimates even if memory space is small.
Keywords: selectivity estimation, query processing, wavelet

1. Introduction

Accurate estimates of the result size of geometric queries are curial to several query processing components of spatial database management system (SDBMS). Selectivity is used in cost-based processor as intermediate results. Sophisticated user interfaces also use estimates of result sizes as feedback to users before a query is actually executed. Since it is infeasible to run the entire query to compute the result sizes, estimating selectivity does on the basis of summary data which is generated by approximation of underlying data. To obtain good selectivity, it is important to make summary data reflecting data distribution perfectly. To do so, we need too much memory space but can do because many applications have required it to be retained with small. It is also hard to get good summary data as spatial search space has enlarged more than more since GIS were used in various fields.

As important and well-known issues to estimate spatial selectivity, there are *false-counting and multi-counting*. To settle above problems, many histograms have been proposed in the literature [2, 3, 4]. We found two ideas to get compression effects in these literatures. First, Previous histograms use axis split which split entire space into two subsets called "bucket"[1]. In the case that all objects are placed on some corner, we generate four buckets. In fact, we need just two buckets in retaining because remain buckets have same frequency. That is, we can reduce two buckets. Second, it is a fact that the distribution of frequencies over the input domain does not vary dramatically in spatial data. If many buckets which have same frequency are placed on some of adjacent area, they can be replaced with one bucket.

Motivated by the above reasoning, we propose a MW Histogram combined MinSkew split method and haar wavelet transform. MinSkew split make buckets which have grid cells with similar frequency and wavelet transform replace such cells with one cell. As a result, MW histogram can not handle well the skewed space but also make compressed summary data.

The rest of this paper is organized as follows. In the next section we summarize related work. The proposed structure and algorithm of MW Histogram is presented in section 2. In section 3 we describe the superiority of our technique through comparing with Wavelet and MinSkew. Finally, we draw conclusions and give a future work in Section 4.

2. Related Work

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

In [2], 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 (Cumulative Density) Histogram and Euler Histogram is proposed in [3, 4]. As in the CD Histogram, Euler Histogram also addresses the multiple-count problem.

In [1] Matias et al , as compress histogram, introduce a new type of histograms, called wavelet-based histograms, based upon multidimensional wavelet decomposition. Wavelet decomposition is performed on the underlying data distribution, and most significant wavelet coefficients are chosen to compose the histogram. In other words, the data points are compressed into a set of numbers via a sophisticated multi-resolution transformation. This approach can be extended very naturally to efficiently compress the joint distribution of multiple attribute. We propose a new method, called MW Histogram, applying one of wavelet-based techniques, Haar Wavelet, to estimate selectivity for spatial range query on skewed spatial datasets.

* This work was supported by University IT Research Center Project and KOSEF RRC Project(Cheongju Univ. ICRC) in Korea

3. MW Histogram

Although MinSkew histogram requires additive memory space so that spatial-skew of buckets comes near to zero and axis split turns out useless buckets, it is advantage that the grid cells have similar frequencies in the same bucket. On the other hand, Wavelet-based histogram show good compression but it is difficult to get reasonable selectivity when data distribution is highly skewed or required space is very little. Fortunately, the two histograms can supplement defects of each other. So, we combine MinSkew spatial partitioning and wavelet transform. Our basic idea is that: if each bucket has minimum skew by partitioning, the frequency of grid cells is getting similar to adjacent grids in the same bucket. if so, the number of coefficients retaining is getting smaller by wavelet transform. This fact gives us good summary data even if the required memory size is very small.

1) The Structure of MW Histogram

The structure of MW histogram is composed of a binary tree and a wavelet synopsis in several buckets (Fig 3.1). A binary tree is composed of split nodes and buckets. The split node has split information such as *split axis, split position, spatial skew along split axis, a pointer of left child and right child.* The bucket has the data distribution of wavelet synopsis mapped into spatial domain such as *spatial skew and wavelet synopsis* which is composed of index and coefficient. By partitioning, if it requires *b* buckets, the number of split nodes is *b-1*. Therefore, Let total memory size is *M. M* is below: *Ws* is the number of wavelet coefficients retaining.

$$M = 5(b \quad 1) + b(1+2Ws) \qquad (1)$$

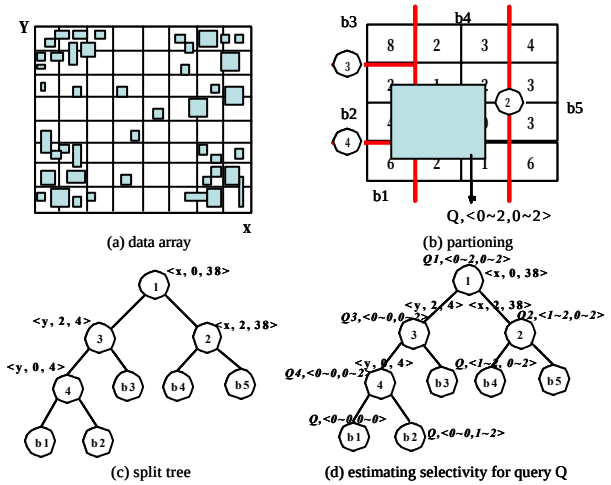

Fig. 1. the structure of MW histogram

2) Construction of MW Histogram

The construction of MW histogram is accomplished by partitioning , wavelet transforming and shrinking step.

Step1. Partitioning

At first, we transform spatial domain to a grid with equal dimension size. The value of each grid cell is allocated by counting all objects intersected with the grid cell and then partitioning is executed. The pseudo-code for the algorithm is below:

Algorithm Partitioning
Start with a single bucket consisting of all the regions
while there are less buckets than needed
  For each current bucket do
    Find the bucket which has the biggest skew along axis among current buckets. The axis of the bucket is split axis.
  EndFor
  Pick the bucket whose split position will lead to the greatest reduction in spatial-skew and then Split the bucket into two and assign regions from the old buckets into the new buckets and generate a split node.
  The split node is arranged into split tree
Endwhile

Step2. Wavelet Transforming

After partitioning, refine the grid cells by splitting each cell into four identical cells and then assign each rectangle in the input to the cells whose MBR contains the center of the rectangle. This step is composed of space-filling ordering and wavelet transforming.

First, Z-mirror, as space-filling order along split axis on each bucket, is executed to transform 2D array into 1D array.

Second, we transform the 1D array into a wavelet synopsis by 1D haar wavelet. Remove then coefficients whose value is zero.

Step3. Shrinking

It assigns the number of retaining coefficients along spatial skew. The higher skew go, the bigger it grows.

3) Selectivity Estimation

Fig. 3.1(d) show selectivity estimates for given query Q <q_{xl}, q_{yl}, q_{xh}, q_{yh}>. Whenever the query visits split nodes, the query is split by split index of the split-nodes along split axis until it reach to buckets. And then, selectivity is computed as sum of estimating input values that is recovered by wavelet recovery function, within a range of each split query.

4) Compression Effects

Bucket of Traditional histogram is composed of six elements. If size of all elements is a unit space, size of one bucket is 6. If total buckets is B, total memory size M is $M=6B$. While total memory size M of MW histogram is like (1) equation if total bucket is b and split nodes is b-1 and Ws is wavelet coefficients in histogram. If both histograms have same memory size, we get an equation as follow:

$$6B = 5(b \quad 1) + b(1+2Ws) \qquad (2)$$

In case that B of MinSkew Histogram is 60 and b is 20, we can compute the number of wavelet coefficients

(Ws) : Ws is 6. However, we consider one coefficient as one bucket because it was mapped into some location in space. In result, MinSkew histogram has 60 buckets same as before but MW histogram has 20*6 = 120 buckets. The fact shows me that MW histogram can obtain compression effect surprisingly. In other word, we can estimate similar selectivity despite a half of memory size of MinSkew.

4. Experimental Evaluation

We compare the effectiveness of MW histogram with MinSkew histogram to lay emphasis of compression effects and reasonable selectivity estimates. We can not find a large amount of spatial data so we use normal data distribution about 11,000 objects. To make similar surrounding, test environment is below:

resolution r : 64, memory size M : 60 ~720 unit
query size $|Q|$: 5% ~ 20%.
MW0~MW2 has a different bucket size as ratio { 0.3, 0.5, 0.7} X M/6
MW-20~40 has a half of memory size M of MinSkew Histogram.

In our experimental result, Fig 4.1~3 show that MW histogram is better on small queries than MinSkew Histogram. Fig 4.1 and Fig 4.3 show that retaining many coefficients help us get good selectivity in very small memory and also MW Histogram maintaining many buckets has low relative error in large memory. Above facts prove that our proposed histogram can usefully apply to spatial database which is on very large spatial domain. In addition, we do not optimize to allocate differently coefficients every bucket. The skewed bucket should have more coefficients than non-skewed buckets. In spite of these facts, our experimental evaluation is quit successful.

5. Conclusions

Previous Histograms require very large memory space to maintain high accuracy of selectivity if spatial domain is also large. Therefore, we proposed a new method called MW histogram that could get reasonable selectivity with small memory size.

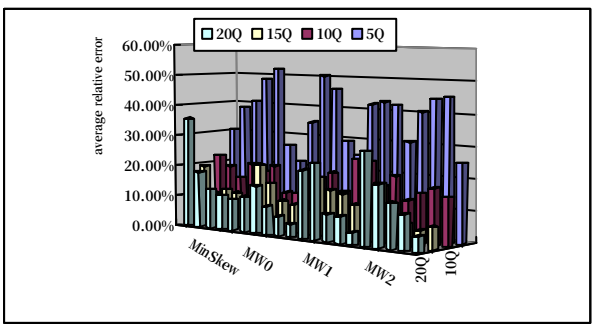

Fig. 2. relative error along query size and memory size

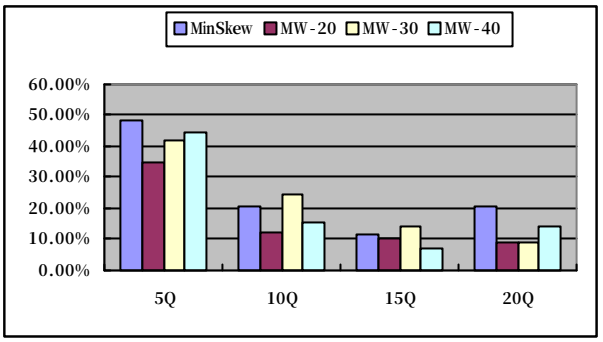

Fig. 3. relative error along query size

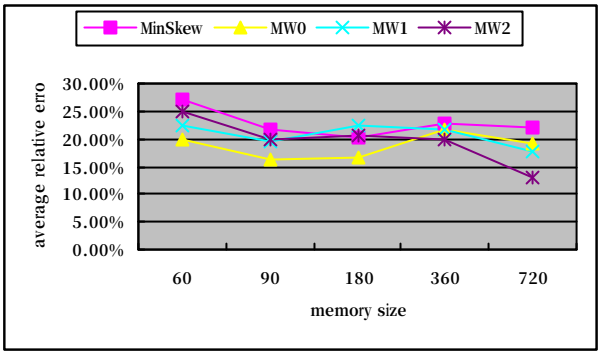

Fig. 4. relative error along memory size

MW histogram combined modified spatial split method with Haar Wavelet transformation so that we obtained maximum compression effects consequently. Based on our experimental and theoretical analysis of the new technique and adaptations of previously known techniques, we are able to show that: (a) MinSkew change selectivity error sensitively by changing memory size. (b) Our technique which called MW Histogram can obtain maximum compression effects and reasonable selectivity simultaneously. Our MW histogram is useful in very large spatial domain.

In the future, we need to analyze our histogram to improve much experimental evaluation. We also will extend our histogram to do work easily about dynamic insertion and updating.

References

[1] 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.

[2] 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.

[3] 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

[4] Sun, C., Agrawal, D., El Abbadi, A., "Exploring spatial datasets with histograms (full version)", Technical Report, Computer Science Department, University of California, santa Barbara, 2001