2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Spatial Selectivity Estimation Using Wavelet

JinYul Lee, JeongHee Chi, KeunHo Ryu
Database Laboratory, Chungbuk National University
Email : { jinylee, jhchi, khryu }@dblab.cbu.ac.kr

Abstract-Selectivity estimation of queries not only provides useful information to the query processing optimization but also may give users with a preview of processing results. In this paper, we investigate the problem of selectivity estimation in the context of a spatial dataset. Although several techniques have been proposed in the literature to estimate spatial query result sizes, most of those techniques still have some drawback in the case that a large amount of memory is required to retain accurate selectivity.

To eliminate the drawback of estimation techniques in previous works, we propose a new method called MW Histogram. Our method is based on two techniques: (a) MinSkew partitioning algorithm that processes skewed spatial datasets efficiently (b) Wavelet transformation which compression effect is proven. We evaluate our method via real datasets. With the experimental result, we prove 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.

I. INTRODUCTION

Recently, as utilization area of geographical information is widely diffused with development of GIS application systems, the interest in management of large geographical information is increasing. Specially, efforts to process complex spatial query efficiently are briskly progressing. We focus on a method of selectivity estimation which is an important component of query optimization.

Spatial selectivity estimation is calculated based on the summary information which is created through approximating spatial data distribution. The summary information approximates the whole space distribution by partitioning it into a suitable number of buckets on retainable memory space. However, it is difficult to exactly approximate spatial data distribution. First of all, most of spatial object distribution in real world has the skewed form like buildings around a river. Morever skewed distribution also contains non-satisfied objects of query during calculating selectivity. We may get the selectivity result with high error. This problem usually is called false counting.

To solve false counting as problem MinSkew split the skewed space along an axis that has largest spatial density into buckets which is a group of areas with equal frequency. They then assign many buckets into skewed area. As a result, a distribution in buckets becomes uniform, so we can get high accurate selectivity. But, the whole space domain become larger, we cannot completely solve the false counting problem as a small number of buckets. To overcome this situation, we use the wavelet transform which is proven to have compression effect on many previous works. We then can retain much summary information in small memory space. In particular, we choose the simplest method, Haar wavelet transform. Haar wavelet has very large compression effect to input sets containing similar values.

After all, in this paper we propose the MW(MinSkew-Wavelet) Histogram. Combined the spatial split algorithm of MinSkew with Wavelet transformation, it can handle well the skewed space distribution and also make the best compressed summary information.

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 III. In section IV we describe the superiority of our technique through comparing with Wavelet and MinSkew. Finally, we draw conclusions and give a future work in Section V.

II. RELATED WORK

Selectivity estimation is a well-studied problem for traditional data types such as integers. 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 [2,4,5,7].

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 is proposed in [7]. Typically when building a histogram for region objects, an object may be counted multiple times if it spans across several buckets. The CD algorithm address this problem by keeping four sub-

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

histogram stores 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 [5]. The mathematical foundation of the Euler Histogram is based on Euler's Formula in graph theory, hence the name Euler Histogram. As in the CD Histogram, Euler Histogram also addresses the multiple-count problem.

Though these techniques are efficient methods to approximate range query selectivity estimation in spatial databases. These techniques require a large amount of memory for better accuracy.

To compress the summary information in conventional databases, In [1] Matias et al. 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. Those coefficients constitute the final histogram. 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.

III. OUR PROPOSED TECHNIQUE

MinSkew histogram requires additive memory space to reduce spatial-skew within each bucket because of axis split method of this histogram. For example, if most of objects are located to a left-upper corner in a bucket, and then we can say this bucket is much skewed. To decrease spatial-skew within the bucket, MinSkew should split the bucket into four buckets. However, we need just one bucket on left-upper side, and rest three buckets can be replaced with one another.

Wavelet histogram has a drawback that compression effect is lower when data distribution is much skewed. In order to hold reasonable selectivity even if it is in highly skewed distribution, the number of coefficients retained is increased. Fortunately, it need to less memory size than axis split method because a coefficient is mapped into a bucket in space.

In order to supplement these histograms mutually, we combine MinSkew spatial split algorithm with wavelet transformation. The basic ideas of our histogram are as follow: if split each bucket has minimum skew, the frequency of each grid cell is similar to one of adjacent grids in each bucket. If so, the number of coefficients which should be retain is getting smaller when wavelet transformation applies to each bucket. The facts give us many information of spatial distribution even if the size of summary information is small.

3.1 The structure of MW histogram

The structure of MW histogram is a binary tree which is composed of spatial split nodes and buckets (figure 3.1). The spatial split node is generated when area of a bucket is split into several buckets in order to reduce spatial-skew of them as small as possible. If the number of retained buckets is *b*, we should split spatial domain *b-1* times. Leaf node then generates two buckets, left bucket and right bucket along split axis of the node. Grid cells of each bucket are sorted by space-filling curve to perform 1-dimensional Haar wavelet transformation in each bucket. At this time, we are faced with a problem for wavelet transformation. If split algorithm of MinSkew applies without any modification, the number of grid cells in some bucket is not satisfied with a condition that perfect wavelet transformation needs 2^n input data. So, we can not perform transformation of 1-dimension Haar wavelet perfectly. Therefore, we modify split algorithm of MinSkew to split according to split-ratio which we define.

Next, we describe construction process of MW histogram as follow.

1. To keep on a condition that the number of grid cells in each bucket should be 2^n, we split buckets according to split-ratio that is choosen by minimizing the sum of spatial-skews of the split buckets.
   Split_Ratio *= { 1:1 ,1:2:1 ,1:1:2 , 2:1:1}*
   split node *< split axis, split index, spatial skew along axis, a pointer of left child, a pointer of right child>*
2. Spatial split is performed *b-1* times to generate *b* buckets. Each leaf-node has two buckets.
3. When all split is end, we perform Haar wavelet transform on each bucket after it is sorted by space-filling cure, Z-ordering, Hilbert ordering, Z-mirror and so on.
   Bucket *< skew of a bucket, wavelet synopsis ∋ {coefficient index, coefficient}>*

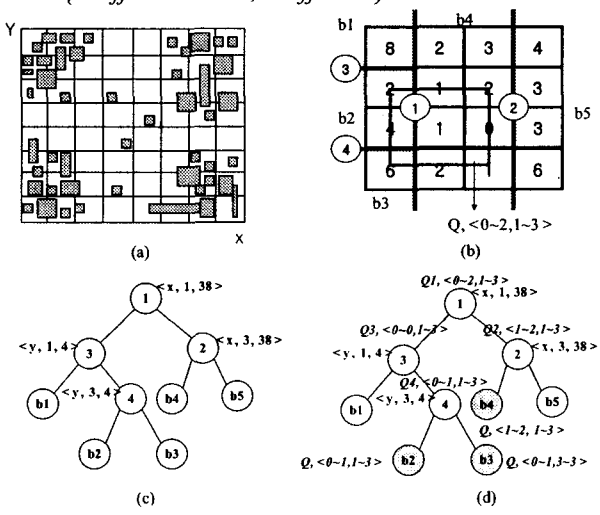

Figure 3.1 the structure of MW histogram

figure 3.1 show MW histogram structure and a process of selectivity estimates. The number of total bucket is 5 and entire input space (a) is split 4 times (c). (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 original input values that is recovered by wavelet recovery function, within a range of each query split.

We describe two algorithms to construct MW histogram at figure 3.2 and figure 3.3

```
ALGORITHM MW_HST(β, Sm, array)
// A number of array is 2^n x 2^m , a number of bucket to retain
in memory β. restricted memory space Sm , Split tree which
is Binary tree BT.
Calculate a number of rectangles intersected with each grid

WHILE β > 0 DO
FOR Bi ∈ LB DO
  Calculate sum of spatial density along each axis
  And then Compute spatial skew of each axis
  IF( Bi.x_skew > Bi.y_skew) THEN
      Max_axis = x, Max_Skew = Bi.x_skew
  ELSE
      Max_axis = y, Max_Skew = Bi.y_skew
  END IF
  IF (Max_B.skew < Max_skew) THEN
      Max_B = Bi, Max_B.axis = Max_axis
  END IF
END FOR
LB. add( FINDSPLIT(Max_B, BT))
LB. delete(Max_B)
β = β – LB.Size
END WHILE
// Wavelet Transformation
Sort buckets in LB order by big skew.
FOR (Bi ∈ LB) DO
   WAVELETTRANSFORM( Bi)
END FOR
END ALGORITHM
```

Figure 3.2 construction algorithm of MW histogram

3.2 Compression Effects

Bucket of Traditional histogram is composed of *<left-bottom point, right-upper point, skew, frequency>*. If size of each element is a unit space, size of one bucket is six. If total buckets is B, total memory size M is $M=6B$. While Total memory size M of MW histogram is $M = 5(b-1) + b(1+2Ws)$, 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: $6B = 5(b-1) + b(1+2Ws)$.

In case that B of MinSkew Histogram is *60* and b is *20*, we can 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 say 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.

```
ALGORITHM FINDSPLIT( Max_B , BT)
// Max_B has a information of a Bucket and a split axis that
spatial skew is the biggest.
// Optimal_Ratio is optimal split ratio which can be minimum
spatial skew.
Split_Ratio = { 1:1, 1:2:1, 1:1:2, 2:1:1 }
MM_Skew = Max_B.skew  // Minimum sum of spatial skew
of each bucket split by split_ratio
// find minimum spatial skew ratio
FOR i = 0,...4 DO
// if each bucket split has bigger spatial skew value than
Max_B, loop will be next.
   // sum of spatial skew of each bucket split by split_ratio.
   Skew = Spatial_Skew(Split_Ratio(i))
   IF  MM_Skew > Skew THEN
       Optimal_Ratio = Split_Ratio(i)
       MM_Skew = Spatial_Skew(Split_Ratio(i))
   END IF
END FOR
  Split(Split_Ratio(i), Max_B) // Max_B is split as selected
split ratio.
  BT.insert(split_nodes, split_buckets)
  RETURN Split_bucket
END ALGORITHM
```

Figure 3.3 split bucket algorithm

IV. EXPERIMENTAL EVALUATION

We compare the effectiveness of MW histogram with MinSkew histogram and Wavelet 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, usable memory size is strictly bounded to be very small. Memory unit is defined as unsigned float type and all elements used in buckets, coefficient, split index and so on, are allocated as same type. It is able to easily compare compression effects of MW histogram with others. To show compression effects and reasonable selectivity estimation of our histogram, we evaluate change ratio of selectivity estimates by a variety of memory size and estimate selectivity by various query size. A range of memory size is from 50 to 600 units and a range of changeable size of query is from two to one hundred times of average object's area.

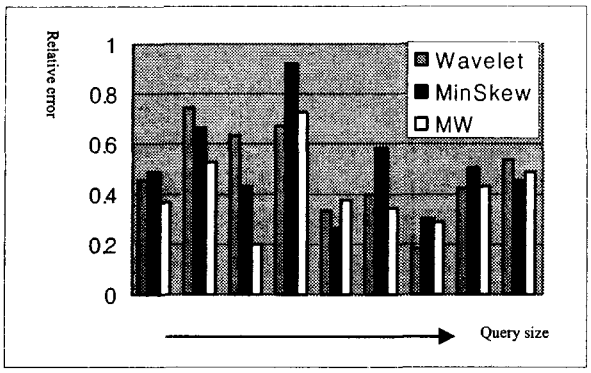

Figure 4.1 relative errors by various query size

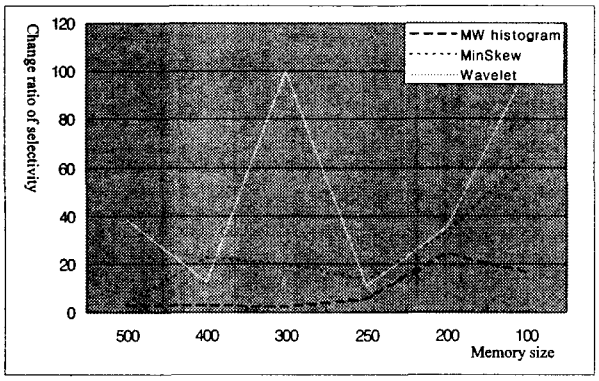

Figure 4.2 change ratio of selectivity by memory size

In our experimental result, MinSkew histogram has normal change ratio when memory size is sufficient to solve spatial skew but high change ratio if not so. Also, Wavelet histogram has high change ratio continuingly. It is said to us the fact that coefficients with large value occur high error when those is removed. In other word, we can say that selectivity estimation using wavelet histogram is bad on distribution of high skew. MW histogram has low change ratio in spite of very small memory size less than 100 units. Figure 4.2 shows that facts are proven. Figure 4.1 depicts that MW histogram get reasonable selectivity. In particular, MW histogram is better on small queries than others. 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.

V. CONCLUSIONS AND FUTURE WORK

Recently, many works have been progressing to not only efficient query processing but also reduction of access costs. Selectivity estimation, one of these works, is used in query optimization and decision of optimal access path cardinally. Until now, several techniques of spatial selectivity estimation have been proposed. These techniques are focused on obtaining high accuracy and fast response time. However, they 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. MW histogram combined modified spatial split method with Haar Wavelet transformation so that we obtained maximum compression effects consequently. In theory, we could save memory costs nearly two times than other histograms. Based on our experimental analysis of the new technique and adaptations of previously known techniques, we are able to show that : (a) MinSkew and Wavelet histogram 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] L. Getoor, B. Taskar, D. Koller, "Selectivity estimation using probabilistic models", In Proc. ACM SIGMOD Int. Conf. on Management of Data, 2001

[4] C. Sun, D. Agrawal, A. El Abbadi, "Selectivity for spatial joins with geometric selections", Proc. of EDBT, 2002, pp.609-626

[5] 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

[6] Minos G., Phillip B.G., " Wavelet Synopses with Error Guarantees", ACM SIGMOD 2002, Jine 4-5, Madison, Wisconsin, USA.

[7] 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

[8] Ning An, Zhen-Yu Yang, Sivasubramaniam A., "Selectivity estimation for spatial joins", In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2001, oo.175-196