일반화된 누적밀도 히스토그램을 이용한 공간 선택율 추정

지 정 희[†] · 김 상 호[†] · 류 근 호[††]

요 약

누적밀도 히스토그램은 사각형 객체의 네 점에 대응하는 4개의 서브 히스토그램을 유지함으로써 사각형 객체가 여러 버켓에 걸쳐질 경우 발생하는 다중 계산 문제를 해결하고 있다. 이 기법은 빠른 추정시간과 정확한 결과를 제공하고 있지만, 질의 윈도우가 그리드 셀의 경계와 일치해야 한다는 제약사항을 기반으로 수행하므로, 실제 응용에 적용시 많은 에러를 초래하게 된다. 따라서, 이 논문에서는 기존 누적밀도 히스토그램에서 질의 윈도우의 제약사항에 관한 영향을 줄이기 위해, 두가지 확률모델을 기반으로 일반화된 누적밀도 히스토그램을 사용한 선택율 추정 기법을 제안하였다. 제안된 두가지 확률 모델은 ① 질의 영역 비율을 고려한 확률모델과, ② 교차 영역 정보를 고려한 확률모델이다. 우리는 실제 데이터 셋을 사용하여 제안된 기법을 실험하였다. 실험 결과는 이 논문에서 제안된 기법이 기존의 다른 선택율 추정 기법보다 성능이 뛰어남을 보여주고 있다. 더구나, 교차 영역 정보를 기반으로 하는 확률모델의 경우 20% 질의 윈도우에서 5% 미만의 낮은 에러율을 보였다. 이 논문에서 제안된 기법은 사각형 객체의 공간 범위 질의의 선택율을 정확하게 추정하는데 사용될 수 있다.

Selectivity Estimation using the Generalized Cumulative Density Histogram

Jeong Hee Chi[†] · Sang Ho Kim[†] · Keun Ho Ryu[††]

ABSTRACT

Multiple-count problem is occurred when rectangle objects span across several buckets. The CD histogram is a technique which solves this 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 propose 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. Our method has the capability of eliminating an impact of the restriction on query window which the existing cumulative density histogram has. We experimented with real datasets to evaluate the proposed methods. Experimental results show that the proposed technique is superior to the existing selectivity estimation techniques. Furthermore, selectivity estimation technique based on probabilistic model considering the intersection area is very accurate(less than 5% errors) at 20% query window. The proposed techniques can be used to accurately quantify the selectivity of the spatial range query on rectangle objects.

키워드 : 공간 질의 최적화(Spatial Query Optimization), 공간 히스토그램(Spatial Histogram), 선택률 추정(Selectivity Estimation), 히스토그램의 일반화(Generalization of Histogram)

1. 서 론

최근 데이터베이스의 크기가 급격히 증가하면서 데이터베이스 시스템의 많은 모듈들이 질의결과 크기의 정확한 추정을 요구하고 있으며, 추정된 질의결과는 다양한 용도로 사용되고 있다. 예를 들면, 질의 최적화기는 대규모 데이터베이스 질의에 대한 효율적인 실행계획을 선택하기 위해 전체 데이터베이스를 접근하는 비효율적인 방식대신, 요약 정보를 기반으로 수행된 선택율 추정결과를 사용하여 최적의 실행계획을 선택하고 있다. 그리고, 질의가 실제적으로 수행되기 전에 질의의 실행 시간에 관한 정보를 사용자에게 제공하기 위해서도 추정된 질의결과를 사용하고 있다. 또한, 데이터웨어하우스에 대한 정확한 응답을 얻기 위해서는 많은 시간과 공간이 필요하므로, 빠른 근사 응답으로 OLAP 질의에 대처하기 위해서도 선택율 추정이 사용되고 있다[3, 12].

공간 데이터베이스에서는 기존 관계형 데이터베이스에서와 같이 실제 데이터 분포를 근사하는 요약 정보를 구성하고 이를 이용하여 선택율을 추정하고 있다[5, 7, 9, 12, 13]. 이러한 요약 정보를 구성하기 위해 제안된 기법들은 버켓 분할 방식과 각 버켓에서 유지되는 정보에 따라 다양한 성능을 보이고 있다. 이 논문에서는 영역 객체의 범위 질의에

※ 이 연구는 2003년도 건교부 국가 GIS사업(국토연구원) 및 대학 IT 연구센터 육성, 지원 사업의 연구비지원으로 수행되었음.
† 준 회 원 : 충북대학교 대학원 전자계산학과
†† 종신회원 : 충북대학교 전기전자 컴퓨터공학부 교수
논문접수 : 2003년 9월 22일, 심사완료 : 2004년 3월 17일

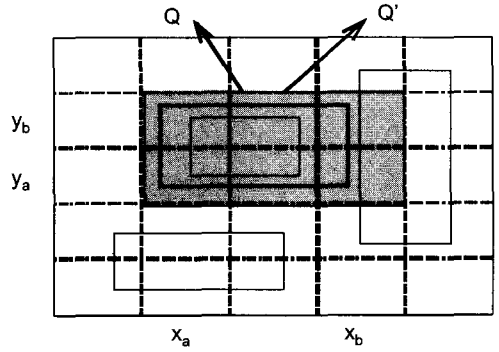
관한 선택율 추정에 초점을 맞춘다. 공간 범위 질의에 대한 선택율 추정기법에 관한 연구로는 [1, 3, 4, 10, 11]가 있고, 특히 영역 객체를 대상으로 하는 히스토그램 기반 접근 방법에는 [3, 4, 7, 9, 11, 13]이 있다. [4, 7]에서는 공간 편중을 줄이기 위한 버켓 분할 기법인 MinSkew 알고리즘과 Quad-Tree 기반 분할 알고리즘을 제안하였고, [3, 11]에서는 웨이블릿(wavelet) 기법을 이용한 히스토그램을 제안하였다. 그렇지만, 이들 기법들은 영역 객체를 포인트 객체로 변환하여 처리하므로, 변환에 필요한 시간 오버헤드를 가지게 된다. [9, 13]에서는 영역 객체의 다중 계산(multiple counting) 문제를 해결하기 위한 누적밀도 히스토그램(cumulative density histogram)과 오일러 히스토그램(Euler histogram)을 제안하였다. 그렇지만, 이들 기법들은 질의 윈도우가 각 그리드 셀의 경계와 일치해야 한다는 제약을 기반으로 수행된다. 선택율 추정을 위한 히스토그램은 가능한 데이터 셋이나 질의 윈도우에 대한 어떠한 가정도 하지 않고, 데이터 셋 자체로부터 정보를 추출할 수 있는 형태로 유지되어야 하지만, 지금까지 제안된 많은 기법들은 데이터 셋이나, 질의 윈도우에 대한 제한된 가정을 기반으로 수행되므로, 실제 응용에 적용될 경우 많은 에러를 초래하게 된다.

따라서, 이 논문에서는 포인트 객체뿐만 아니라, 사각형 객체의 범위 질의에 관한 정확한 선택율을 추정하는 누적밀도 히스토그램의 질의 윈도우에 대한 제약사항의 영향을 최소화하기 위해 두 가지 확률모델을 기반으로 히스토그램의 일반화를 수행하여 선택율 추정하는 기법을 제안한다. 첫번째는 기존의 MinSkew 히스토그램에서 적용하고 있는 확률모델로 데이터 영역과 질의 윈도우 교차 영역의 비율을 고려한 모델이며, 두번째는 데이터의 교차 영역 정보를 이용한 확률모델이다. 실험을 통해서 이 논문에서 제안한 교차 영역 정보를 이용한 선택율 추정기법이 적은 공간 오버헤드를 초래하지만, 기존에 제안된 다른 기법들에 비해 낮은 에러율을 보이고, 선택율 추정을 빠르게 한다는 사실을 보여주었다.

이를 위한 논문의 구성은 다음과 같다. 먼저 2장에서는 문제 정의 및 연구 동기에 대하여 기술하고, 3장에서는 이 논문의 연구 기반이 되는 누적 히스토그램과 공간 범위 질의에 관한 기존 연구를 고찰한다. 다음으로 4장에서는 두가지 확률모델을 통한 누적 히스토그램의 일반화 기법을 제시하고 5장에서는 실험 및 성능 평가를 수행한 후 6장에서 결론 및 향후 연구 방향을 제시한다.

2. 문제 정의 및 연구 동기

누적밀도 히스토그램은 사각형 객체의 다중 계산 문제를 효율적으로 처리하여 적은 공간 오버헤드를 가지고 빠르고, 높은 선택도를 유지하는 기법이다[9]. 또한, 이 기법은 영역 객체의 MBR을 기반으로 히스토그램을 구성하므로, 별도의 요약정보를 구성하기 위해 데이터에 접근하지 않아도 되며, 인덱스를 생성할 때 동시에 생성할 수 있다. 그러나, 질의 윈도우가 주어진 그리드 셀의 경계와 일치하여야 한다는 제약 사항을 기반으로 수행되므로, 실제 응용에 적용시 많은 에러를 발생하게 된다.



(그림 1) 누적밀도 히스토그램의 질의 예

즉, (그림 1)의 예에서처럼 질의 Q가 주어졌을 때, 누적밀도 히스토그램은 셀의 경계와 일치하는 Q'과 같은 질의 윈도우를 사용하여 선택율을 추정한다. 따라서, 실제 질의 Q의 선택율은 1임에도 불구하고, 질의 Q'에 관한 선택율은 2를 추정하게 되어, 100%의 에러율을 나타내게 된다.

따라서, 누적밀도 히스토그램을 실제 응용에 적용시키기 위해서는 질의 윈도우에 대한 제약사항의 영향을 줄일수 있는 기법이 필요하다. 이 논문에서는 히스토그램은 데이터 셋이나 질의 윈도우에 대한 제약사항이 적어야 실제 응용에 적용 가능하다는 연구 동기를 기반으로 확률모델을 이용한 누적밀도 히스토그램의 일반화를 통한 선택율 추정기법을 제시한다.

3. 기존 공간 선택율 추정 및 누적밀도 히스토그램

이 절에서는 공간 선택율 추정 기법에 살펴보고, 이 논문의 기반이 되는 누적밀도 히스토그램의 특성에 대해 기술한다.

3.1 공간 선택율 추정

공간 선택율은 주어진 공간 질의와 교차하는 공간 객체의 수를 의미한다. 공간 데이터베이스에서는 파라메트릭 기법, 샘플링, 히스토그램을 기반으로 선택율을 추정하고 있다. 파라메트릭 기법은 데이터 셋에 대한 특성을 단순화한 기법으로 상당히 제한적이며, 데이터 셋의 분포에 의존적이다[1, 2, 6, 10]. 샘플링 기법은 샘플이라는 작은 데이터 셋에 대한 질의를 수행함으로써 전체 데이터 셋에 대한 선택율을 추정하는 것으로, 샘플이 잘 선정되었을 때는 우수한 성능을 보이지만, 그렇지 않을 경우는 상당한 에러를 발생한다. 또한, 샘플링된 데이터를 재사용할 수 없다는 단점을 가지고 있다[12]. 히스토그램 기법은 공간 상의 객체들에 대한 정보를 보조적인 데이터 구조에 유지하고, 질의가 주어질

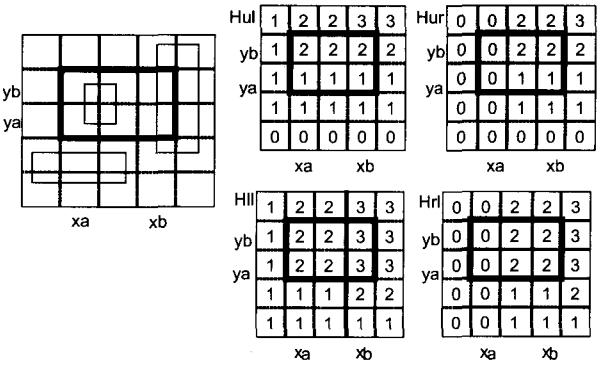
때 선택율을 추정하기 위해 사용한다. 이 기법은 입력데이터에 대한 정보를 요구하지 않고, 작은 공간을 사용하여 효율적으로 유지될 수 있으므로 가장 많이 사용되고 있다[9, 12, 13].

[4]에서는 기존의 히스토그램 방식이 데이터가 균일하게 분포되어 있다는 가정을 기반으로 하기 때문에, 편중된 데이터 분포를 갖는 데이터 셋에 대해서는 높은 에러가 발생한다는 문제점을 해결하기 위해, 공간 데이터를 근사화하기 위한 MinSkew 히스토그램을 제안하였다. 이 기법은 상대적으로 높은 정확도와 적은 저장공간을 사용한다고 제시되고 있다. 또한 [13]에서는 그래프 이론의 오일러 공식을 기반으로 사각형 객체의 다중 계산 문제를 해결하기 위해 오일러 히스토그램을 제안하였다. 이 기법은 R^2의 그리드에 대해, 그리드의 내부 정점, 선분, 면(face)에 대응하는 각 버켓에 히스토그램을 생성한다. 히스토그램의 버켓은 각각의 정점, 선분, 면과 교차하는 객체의 수를 저장한다. 이 기법은 2차원 객체만을 대상으로 하며, 질의 윈도우가 그리드 셀의 경계와 일치해야 한다는 가정을 기반으로 영역 객체의 선택율을 추정하고 있다. [9]에서는 오일러 히스토그램과 같이 영역 객체의 다중 계산 문제를 해결하기 위해 누적밀도 히스토그램을 제안하였다. 이 기법은 포인트 객체와 사각형 객체 모두에 적용가능하며, 상대적으로 정확한 선택도를 유지한다. 이 기법은 이 논문의 기반이 되는 기법으로 3.2절에서 자세히 설명한다. [11]에서는 신호처리와 영상처리에서 사용되는 웨이블릿 변환(Wavelet Transformation)을 이용하여 웨이블릿 히스토그램을 제안하였다. 이 기법은 영역 객체를 포인트 객체로 변환후 웨이블릿 기법을 적용하여 적은 저장 공간을 사용하여 비교적 높은 선택도를 유지한다.

3.2 누적밀도 히스토그램

누적밀도 히스토그램은 사각형 객체의 범위 질의 선택율을 근사처리하기 위해 제안된 기법이다[15]. 이 기법은 모든 객체들을 그들의 MBR에 의해 표현하며, 따라서, 인덱스를 생성할 때 동시에 요약정보를 생성할 수 있다. 누적밀도 히스토그램은 2차원 R^2의 그리드가 주어지면, 4개의 서브히스토그램(H_{ll}, H_{lr}, H_{ul}, H_{ur})을 생성한다. 각 히스토그램 크기는 각각의 그리드 셀에 대응하는 버켓과 동일한 크기 N을 갖는다. H_{ll} 버켓은 버켓에 할당된 왼쪽하단 정점(Lower-Left Vertices)의 수를 유지하며, H_{lr}, H_{ul}, H_{rl}은 오른쪽 하단 정점(Lower-Rightt Vertices), 왼쪽상위 정점(Upper-Left Vertices), 오른쪽상위 정점의 수를 유지한다. 질의 효율성을 향상시키기 위해, 모든 히스토그램은 누적된다. 버켓 H(i, j)는 영역(0, 0, i, j)의 정점수를 저장한다. 따라서 질의 Q(xa, ya, xb, yb)에 교차하는 객체의 수는 다음과 같이 계산될 수 있다.

$$N_{intersect} = H_{ll}(x_b, y_b) - H_{lr}(x_{a-1}, y_b) - H_{ul}(x_b, y_{b-1}) + H_{ur}(x_{a-1}, y_{a-1})$$



(a) 데이터 셋과 질의 윈도우 (b) 누적밀도 히스토그램

(그림 2) 누적밀도 히스토그램의 예

(그림 2)는 누적밀도 히스토그램의 예를 보여주고 있다. (그림 2-a)와 같은 데이터 셋과 질의 윈도우 Q(xa, ya, xb, yb)가 주어지면, 선택율은 $N_{intersect} = 3 - 0 - 1 + 0 = 2$로 추정된다. 누적밀도 히스토그램은 임의의 질의와 교차하는 객체의 수를 정확하게 반환하며, R^d의 그리드에 대한 저장장소는 4^h 요구하며, 질의에 답하기 걸리는 시간은 2^d이므로, 상수시간에 답할 수 있다.

4. 누적밀도 히스토그램의 일반화

이 절에서는 먼저 우리가 제안하는 누적밀도 히스토그램의 일반화를 수행하기 위해 사용되는 용어 및 파라메터들에 대해 기술하고, 다음으로 두가지 확률모델을 이용한 누적밀도 히스토그램의 일반화를 제안한다.

4.1 용어 정의

공간 히스토그램은 가능한 데이터 셋이나 질의 윈도우에 대한 제한적 가정을 갖지 않아야 하며, 따라서 실제 응용에 기존 기법을 적용하기 위해서는 히스토그램을 일반화할 필요가 있다. 이 절에서는 일반화된 누적밀도 히스토그램을 이용하여 선택율을 추정하기 위해 사용되는 용어를 정의한다. <표 1>은 일반화된 누적밀도 히스토그램에서 사용되는 파라메터들을 나타내고 있다.

〈표 1〉 일반화된 누적밀도 히스토그램을 위한 파라메터

기 호	기 호 의 미
Q	(xa, ya, xb, yb) 좌표를 갖는 질의 윈도우
q_i	$q_i \in Q$, i = 1, 2, ···, n
B_Q	질의 Q와 교차하는 버켓
S(Q)	질의 Q에 대한 실제 선택율
S(Q)	질의 Q에 대한 추정된 선택율
Hll(I, j)	셀 (I, j)까지 누적된 왼쪽 하단에 관한 정점의 수
Hlr(I, j)	셀 (I, j)까지 누적된 오른쪽 하단에 관한 정점의 수
Hul(I, j)	셀 (I, j)까지 누적된 왼쪽 상단에 관한 정점의 수
Hur(I, j)	셀 (I, j)까지 누적된 오른쪽 상단에 관한 정점의 수
Area(Q)	질의 Q의 면적
Area(B_Q)	질의 Q와 교차하는 그리드 셀의 면적
xWidth	그리드 셀의 폭
yHeight	그리드 셀의 높이
AE	평균 상대적 에러(Average Relative Error)

[정의 1] 누적밀도 히스토그램

영역 객체의 MBR에 의해 요약 정보를 구성하며, 그리드 셀의 수와 같은 N개의 버켓을 사용하고, 각 버켓에는 H_{ll}, H_{lr}, H_{ul}, H_{ur} 정보를 저장한다. 즉, 누적밀도 히스토그램 CDH는 다음과 같이 나타낸다.

$$CDH(i, j) = \{ H_{ll}(i, j),\ H_{lr}(i, j),\ H_{ul}(i, j),\ H_{ur}(i, j)\} \quad \square$$

여기서, 누적밀도 히스토그램은 그리드 레벨 h가 주어지면 4^h개의 셀이 생성하고, 4^h개와 동일한 수의 버켓을 할당한다. i, j는 버켓의 위치를 나타낸다. 각 버켓에는 H_{ll}, H_{lr}, H_{ul}, H_{ur} 정보를 저장하며, 이들 정보에 대한 내용은 다음과 같다.

- $H_{ll}(I, j)$은 그리드셀 (I, j)에 속하는 사각형 객체의 왼쪽 하단 코너점의 수를 누적하여 저장하며, 식 (1)과 같이 계산된다.

$$H_{ll}(i, j) = \sum_{x=0}^{x=j} BS(i, x) \tag{1}$$

여기서 BS(I, j)는 (0, j)에서 (I, j)까지 범위에 놓여 있는 왼쪽 하단 코너점을 갖는 객체의 수를 저장하고 있다. 식 (1)에서와 같이 $H_{ll}(I, j)$는 결국 (0, 0)에서 (I, j)까지의 범위에서 왼쪽 하단 코너점을 갖는 객체의 수를 저장하게 된다.

- $H_{lr}(I, j)$은 그리드셀 (I, j)에 속하는 사각형 객체의 오른쪽 하단 코너점의 수를 누적하여 저장하며, 식 (2)과 같이 계산된다.

$$H_{lr}(i, j) = \sum_{x=0}^{x=j} BE(i, x) \tag{2}$$

여기서 BE(I, j)는 (0, j)에서 (I, j)까지의 범위에 놓여 있는 오른쪽 하단 코너점을 갖는 객체의 수를 저장하고 있다. 식 (2)에서와 같이 $H_{lr}(I, j)$은 결국 (0, 0)에서 (I, j)까지의 범위에서 오른쪽 하단 코너점을 갖는 객체의 수를 저장하게 된다.

- H_{ul} (I, j)은 그리드셀 (I, j)에 속하는 사각형 객체의 왼쪽 상단 코너점의 수를 누적하여 저장하며, 식 (3)과 같이 계산된다.

$$H_{ul}(i, j) = \sum_{x=0}^{x=j} US(i, x) \tag{3}$$

여기서 US(I, j)는 (0, j)에서 (I, j)까지 범위에 놓여 있는 왼쪽 상단 코너점을 갖는 객체의 수를 저장하고 있다. 식 (3)에서와 같이 $H_{ul}(I, j)$은 결국 (0, 0)에서 (I, j)까지의 범위에서 왼쪽 상단 코너점을 갖는 객체의 수를 저장하게 된다.

- H_{ur} (I, j)은 그리드셀 (I, j)에 속하는 사각형 객체의 오른쪽 상단 코너점의 수를 누적하여 저장하며, 식 (4)과 같이 계산된다.

$$H_{ur}(i, j) = \sum_{x=0}^{x=j} UE(i, x) \tag{4}$$

여기서 UE(I, j)는 (0, j)에서 (I, j)까지 범위에 놓여 있는 오른쪽 상단 코너점을 갖는 객체의 수를 저장하고 있다. 식 (4)에서와 같이 $H_{ur}(I, j)$은 결국 (0, 0)에서 (I, j)까지의 범위에서 오른쪽 상단 코너점을 갖는 객체의 수를 저장하게 된다.

위의 내용과 같이 생성된 요약정보를 기반으로 선택율을 추정한다. 선택율 추정은 주어진 질의 윈도우 Q와 교차하는 객체의 수를 추정하는 것이며, 누적밀도 히스토그램에서는 다음과 같은 식을 이용하여 추정한다.

$$S'(Q) = (H_{ll}(xb, yb) - H_{lr}(xa-1, yb) - H_{ul}(xb, yb-1) + H_{ur}(xa-1, ya-1)) * P \tag{5}$$

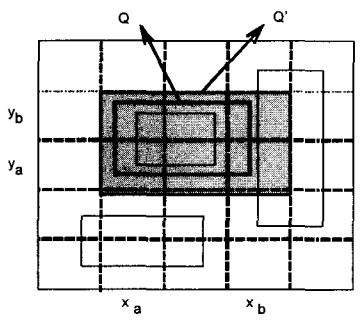
여기서, P는 히스토그램을 일반화시키기 위한 확률모델을 의미하며, 다양한 확률모델을 결합하여 선택율의 정확도를 높일 수 있다. 이 논문에서 제안한 확률모델에 관해서는 4.2와 4.3절에서 설명한다. 또한, 식 (5)에 의해 추정된 선택율의 정확도는 평균 상대적 에러율에 의해 측정하며, 평균 상대적 에러율은 다음과 같은 식으로 표현된다.

$$AE = \frac{(\sum_{qi \in Q} |S'(qi) - S(qi)|)}{\sum_{qi \in Q} S(qi)} * 100 \tag{6}$$

여기서, qi는 i번째 질의 윈도우를 나타내며, 이 논문에서는 동적으로 생성한 20개의 질의 윈도우를 기반으로 실험하였다.

4.2 질의 영역 비율을 이용한 히스토그램의 일반화

히스토그램을 실제 응용에 적용하기 위해서는 데이터 셋과 질의 윈도우에 대한 제약사항이 적어야 한다. 이 절에서는 누적밀도 히스토그램의 질의 윈도우에 대한 제약사항의 영향을 줄이기 위하여 MinSkew 히스토그램 등의 기법에서 사용되고 있는 질의 영역 비율을 기반으로 하는 확률모델을 이용하여 누적밀도 히스토그램의 일반화를 수행한다.



(그림 3) 누적밀도 히스토그램의 질의 윈도우 Q 와 Q'

(그림 3)은 질의 윈도우 Q와 누적밀도 히스토그램을 이용하기 위해 확장된 질의 Q'을 보여주고 있다. 질의 영역 비율을 이용한 확률모델은 질의 Q와 질의와 교차하는 영역

비율을 의미하며, 다음과 같은 식에 의해 구해질 수 있다.

$$P = \frac{Area(Q)}{Area(B_Q)} \tag{7}$$

여기서, Area(Q)는 질의의 면적을 나타내며 (그림 3)에서의 질의 Q의 영역을 나타낸다. 이 영역은 질의의 x축 길이와 y축 길이를 이용하여 구해질 수 있으며, 식 (8)과 같이 계산될 수 있다. 또한 Area(B_Q)는 질의 Q와 교차하는 그리드 셀의 면적을 내며 즉, (그림 3)에서의 확장된 질의 Q'의 넓이와 같게 된다. 확장된 질의 Q'의 넓이는 그리드 셀과 교차하는 x, y 그리드 셀의 수와 그리드 셀의 x, y 길이를 이용하여 구할 수 있으며, 식 (9)에 의해 계산될 수 있다.

$$Area(Q) = (xb - xa) * (yb - ya) \tag{8}$$

$$Area(B_Q) = ((H_{xll} - H_{xur} + 1) * xWidth) * ((H_{yll} - H_{yur} + 1) * yHeight) \tag{9}$$

여기서, H_{xll}, H_{xur}은 각각 H_{ll}과 H_{ur}의 x축 그리드 셀의 위치를 의미하며, H_{yll}, H_{yur}은 각각 H_{ll}과 H_{ur}의 y축 그리드 셀의 위치를 의미한다.

따라서, 식 (7)과 같이 정의된 확률 P를 이용하여 누적밀도 히스토그램의 선택율을 추정하기 위한 식 (5)는 다음과 같이 재작성될 수 있다.

$$s'(Q) = (H_{ll}(xb, yb) - H_{lr}(xa-1, yb) - H_{ul}(xb, yb-1) + H_{ur}(xa-1, ya-1)) * \frac{Area(Q)}{Area(B_Q)} \tag{10}$$

간단한 예를 들면, (그림 2)와 (그림 3)의 예에서, 질의 Q가 (1.1, 2.3, 3.4, 3.7)와 같은 좌표를 갖는다고 가정하면 선택율은 S'(Q)= (3-0-1+0) * (3.22/6) = 1.07과 같이 추정된다. 따라서, 질의 영역 비율 기반의 확률모델을 이용하여 히스토그램을 일반화 시킬 경우 정확한 선택율을 추정할 수 있다.

4.3 교차 영역 정보를 이용한 히스토그램 일반화

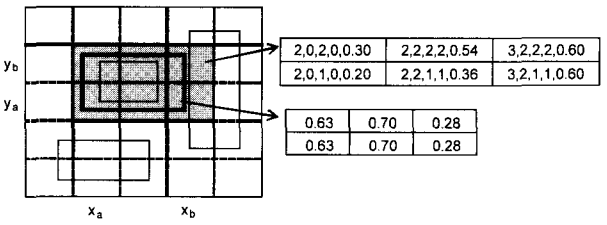
히스토그램은 각 버켓에 유지되는 정보에 따라 다양한 성능을 보인다. 이 절에서는 기존의 누적밀도 히스토그램에 각 셀과 교차되는 정보를 추가적으로 유지함으로써 교차하는 영역 정보를 기반으로 하는 확률모델을 이용하여 누적밀도 히스토그램의 일반화를 수행한다.

교차 영역에 관한 정보를 고려하기 위해서 누적밀도 히스토그램은 다음과 같이 재정의 되며, (그림 4)는 교차 영역 정보를 포함하는 누적밀도 히스토그램의 예를 보여주고 있다.

[정의 2] 교차 영역정보를 고려한 누적밀도 히스토그램

영역 객체의 MBR에 의해 요약 정보를 구성하고, 그리드 셀의 수와 같은 N개의 버켓이 필요하다. 각 버켓은 Hll, Hlr, Hul, Hur, iArea 정보를 저장한다. 즉, 교차 영역정보를 고려한 누적밀도 히스토그램 CDH는 다음과 같이 나타낸다.

CDH(i, j) = { H_{ll}(i, j), H_{lr}(i, j), H_{ul}(i, j), H_{ur}(i, j), iArea(I, j)} □



(그림 4) 교차 영역 정보를 포함하는 누적밀도 히스토그램의 예

여기서, iArea 정보는 각 셀과 교차하는 객체의 교차 면적을 의미한다. 이러한 교차 영역 정보를 이용한 확률 모델은 각 셀과 교차하는 객체 영역과 질의 영역의 비율 대비 질의 영역에 속하는 교차 영역의 비율로서 나타낼 수 있다. 즉, 각 셀과 교차하는 객체 영역과 질의 영역은 식 (11)에 의해 구해질 수 있다.

$$\sum_{i=k}^{i=l} \sum_{j=m}^{j=n} iArea(i, j) * Area_{i,j}(Q) \tag{11}$$

여기서, k, l, m, n은 각각 질의와 교차하는 x축과 y축의 범위를 나타내며, iArea(I, j)는 셀(I, j)와 교차하는 객체의 교차면적 비율을 나타낸다. 또한, $Area_{i,j}$(Q)는 셀(I, j)와 교차하는 질의 윈도우의 교차 면적 비율을 나타낸다. 질의 영역에 속하는 교차 영역의 비율은 k, l, m, n의 범위내의 교차 영역의 합을 의미하므로, 그 범위에 속하는 iArea 값의 합으로 나타낼 수 있다. 따라서, 교차 영역 정보를 이용한 확률 모델은 다음과 같은 식 (12)로 나타낼 수 있다.

$$P = \frac{\sum_{i=k}^{i=l} \sum_{j=m}^{j=n} iArea(i, j) * Area_{i,j}(Q)}{\sum_{i=k}^{i=l} \sum_{j=m}^{j=n} iArea(i, j)} \tag{12}$$

식 (12)과 같이 정의된 확률모델을 기반으로 선택율을 추정하기 위한 식 (5)는 다음과 같이 재작성될 수 있다.

$$s'(Q) = (H_{ll}(xb, yb) - H_{lr}(xa-1, yb) - H_{ul}(xb, yb-1) + H_{ur}(xa-1, ya-1)) * \frac{\sum_{i=k}^{i=l} \sum_{j=m}^{j=n} iArea(i, j) * Area_{i,j}(Q)}{\sum_{i=k}^{i=l} \sum_{j=m}^{j=n} iArea(i, j)} \tag{13}$$

(그림 4)의 예에서 교차 영역정보를 기반으로 하는 확률을 이용하여 선택율 S'(Q) = (3-0-1-0) * (1.281/2.6) = 0.99과 같이 추정된다. 따라서, 교차 영역정보 기반의 확률모델을 이용하여 히스토그램을 일반화시킬 경우 정확한 선택율을 추정함을 알 수 있다.

5. 히스토그램의 성능 및 효율성 분석

이 절에서는 일반화된 누적밀도 히스토그램의 효율성을 평가하기 위해, 공간 범위 질의에 성능이 뛰어나다고 제안

되고 있는 MinSkew 히스토그램, 웨이블릿 히스토그램과 비교 평가한다.

5.1 실험 환경

이 논문에서 제안된 기법의 정확도를 측정하기 위해 512MB의 주 메모리, 40G 하드디스크를 가진 윈도우 2000 운영체제하의 Intel Pentium Ⅳ 2GHz PC에서 실험하였다. 실험 데이터 셋으로는 ① 실제 서울 중구의 건물 10,484개(D1 데이터셋), ② TIGER/LINE의 캘리포니아 데이터셋 2,249,727개(D2 데이터셋), ③ Sequioa 2000 벤치마크를 위한 22,288개의 폴리곤 데이터셋(D3 데이터셋)을 대상으로 실험하였다. (그림 5)는 D1 데이터셋의 분포를 보여 주고 있다.

(그림 5) 서울 중구 상업용 건물의 분포

또한, 우리는 서로 다른 질의 윈도우 크기에 대한 영향을 평가하기 위해 전체 영역의 5%, 10%, 15%, 20%의 질의 윈도우 20개를 임의로 생성하여 실험하였다.

5.2 평가 기준

제안된 기법의 성능 및 효율성을 평가하기 위해, 선택율 추정 에러, 선택율 추정시간, 히스토그램 생성 시간 같은 메트릭을 고려하였다.

5.2.1 선택율 추정 에러

실제 범위 질의에 대한 선택율에 따른 퍼센트로서 실제 실행된 범위 질의의 선택율과 이 논문에서 제안된 기법에 의해 추정된 선택율의 차를 나타내며, 식 (6)과 같이 계산된다. 선택율 추정에러는 낮을수록 좋은 성능을 나타낸다.

5.2.2 선택율 추정 시간

각 기법에 따라 선택율을 추정하는데 걸리는 시간을 의미한다. 선택율 추정 시간을 빠를수록 좋은 성능을 나타낸다.

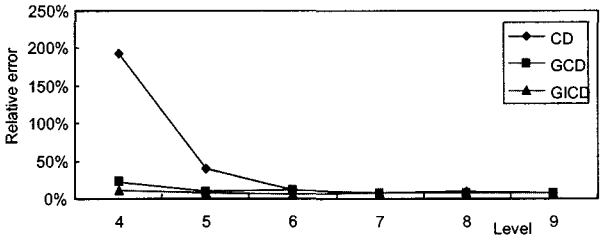
5.2.3 히스토그램 생성 시간

각 기법에 따른 히스토그램의 필요한 정보를 생성하는데 걸리는 시간으로, 생성 시간이 빠를수록 좋은 성능을 나타내지만, 공간 히스토그램의 경우 대부분 오프라인으로 생성되기 때문에, 선택율 추정시간보다는 큰 이슈가 되지 못한다.

5.3 실험 결과

5.3.1 선택율 추정 에러율

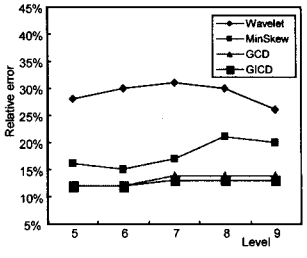
제안된 기법에 의한 추정치의 정확도를 측정하기 위해서 식 (6)의 평균 상대적 에러율을 사용하였다. 임의의 20개 질의에 대한 실험 결과들에 대한 평균을 구하여 에러율을 측정하였다. 먼저, 기존 누적밀도 히스토그램의 질의 윈도우에 대한 제약사항을 무시하고, 응용에 적용했을 때의 에러율과 일반화된 누적밀도 히스토그램의 에러율을 비교 평가 하였다.



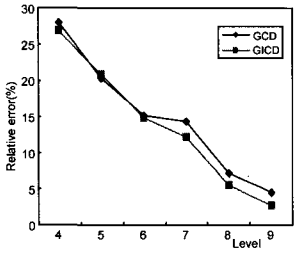
(그림 6) 누적밀도 히스토그램과 일반화된 누적밀도 히스토그램의 평균 상대적 에러율

(그림 6)은 기존의 누적밀도 히스토그램과 일반화된 누적밀도 히스토그램의 평균 상대적 에러율을 보이고 있다. 기존의 누적밀도 히스토그램의 경우 일반화된 누적밀도 히스토그램보다 높은 에러율을 보이고 있다. 그렇지만, 그리드 레벨이 9인 경우에는 일반화된 누적밀도 히스토그램과 비슷한 성능을 보이고 있다. 이러한 이유는 그리드 레벨이 상세화될수록 질의 윈도우와 그리드 레벨의 경계가 일치할 확률이 높아지기 때문이다. 기존의 누적밀도 히스토그램의 경우 일반화된 누적밀도 히스토그램보다 상당히 높은 에러율을 나타내므로, 기존의 다른 선택율 추정 기법과의 비교 시에는 제외하였다.

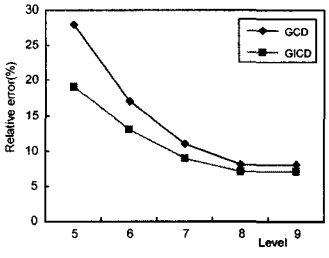
일반화된 히스토그램의 효율성을 평가하기 위해 기존에 제안된 공간 질의 선택율 추정 기법중 그 성능이 우수하다고 제안되고 있는 MinSkew 히스토그램과 웨이블릿 히스토그램과 비교 평가하였다. (그림 7)은 세가지 데이터 셋에 대한 그리드 레벨에 따른 평균 상대적 에러율을 보이고 있다. (그림 7)(a)에서 보여주듯이 이 논문에서 제안된 기법인 질의 영역비율을 이용한 GCD 기법과, 교차 영역정보를 이용한 GICD 기법이 웨이블릿 보다는 약 80%, MinSkew보다는 약 24% 뛰어난 것으로 나타냈다. 특히, (그림 7)(b), (그림 7)(c)에서 보여주듯이 교차 영역 정보를 고려한 GICD 기법은 질의 영역을 고려한 GCD 기법보다 약 4%정도 뛰어난 성능을 보이는 것으로 평가되었다.



(a) D1 데이터 셋

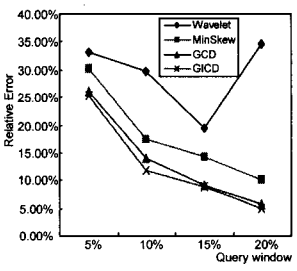


(b) D2 데이터 셋

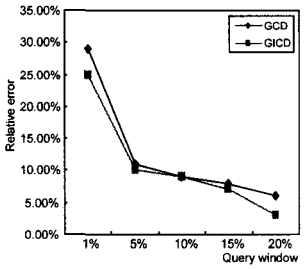


(c) D3 데이터 셋

(그림 7) 그리드 레벨에 따른 평균 상대적 에러율



(a) D1 데이터셋



(b) D3 데이터셋

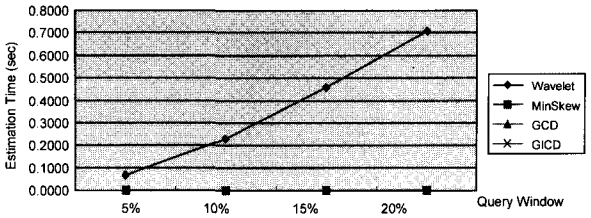
(그림 8) 질의 윈도우 크기에 따른 평균 상대적 에러율

(그림 8)은 질의 윈도우 크기에 따른 에러율의 영향을 평가하기 위해 D1과 D3 데이터셋을 기반으로 수행한 실험결과이다. (그림 8)(a)의 실험결과에서 보여주듯이 웨이블릿 히스토그램을 제외한 MinSkew, GCD, GICD 기법은 질의 윈도우 크기가 증가함에 따라 선택율의 정확도가 증가하는 경향을 보였으며, (그림 8)(a), (그림 8)(b)의 실험결과에서는 GICD 기법의 경우, 질의 윈도우 크기가 20%일 경우에 5% 이하의 비교적 정확한 선택율을 추정함을 보여주었다.

5.3.2 선택율 추정 및 히스토그램 생성 시간

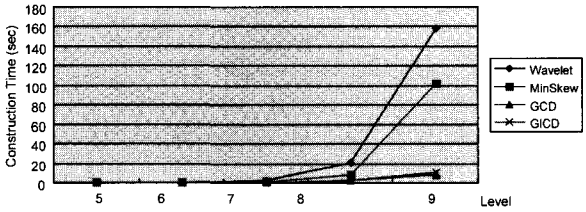
선택율 추정 작업은 오프라인에서 히스토그램을 생성하기 위한 작업을 수행하며, 온라인에서는 생성된 히스토그램을 기반으로 선택율을 추정한다. 일반적으로 히스토그램 생성시간보다는 선택율 추정 시간이 중요한 이슈로 다루어지고 있다.

(그림 9)는 질의 윈도우 크기에 따른 선택율 추정시간을 보여주고 있다. 이 실험은 D1 데이터셋을 기반으로 실험한 결과이다. 실험 결과는 웨이블릿 히스토그램의 경우 질의 윈도우를 1차원 범위 질의로 변환하는 과정이 필요하므로 상당히 많은 시간 오버헤드를 초래한다는 것을 보여주고 있다.



(그림 9) 질의 윈도우 크기에 따른 선택율 추정시간

이에 반해, MinSKew, GCD, GICD는 비교적 낮은 추정 시간을 보였다. 특히, GCD의 경우 상수시간(2^d)에 응답할 수 있다.



(그림 10) 그리드 레벨에 따른 히스토그램 생성시간

(그림 10)은 D1 데이터셋을 기반으로 한 히스토그램의 생성시간을 보여 주고 있다. 웨이블릿과 MinSkew 히스토그램의 경우 레벨에 증가함에 따라 상당히 많은 시간 오버헤드를 초래하였지만, GCD와 GICD 히스토그램의 경우 비교적 낮은 생성 시간을 초래하였다.

지금까지의 실험 결과 이 논문에서 제안된 기법은 기존의 다른 선택율 추정 기법보다 우수한 것으로 평가되었다. 특히, 교차 영역 정보를 기반으로 하는 GICD 히스토그램의 경우 질의 영역을 고려하는 GCD 히스토그램보다는 약 0.26초의 생성시간에 대한 오버헤드를 가지지만, 선택율 추정면에서는 GCD보다 약 4%정도 뛰어난 것으로 평가되었다.

6. 결 론

사각형 객체의 경우 여러 개의 그리드 셀에 걸쳐질 경우 다중 계산되는 문제로 인하여 선택율 추정시 많은 오류를 발생한다. 이러한 문제점을 해결하기 위해 제안된 누적밀도 히스토그램은 빠른 추정시간과 정확한 선택율을 제공한다. 그러나, 질의 윈도우가 그리드 셀의 경계와 일치해야 한다는 제약사항을 기반으로 수행되므로, 실제 응용에 적용시 많은 에러를 초래하게 된다. 따라서, 이러한 문제점을 해결하기 위해 이 논문에서는 두가지 확률모델을 기반으로 누적밀도 히스토그램의 일반화를 수행하였다. 제안된 두가지 확률 모델은 ① 질의 영역 비율을 고려한 확률모델과, ② 교차 영역 정보를 고려한 확률모델이다. 우리는 실제 데이터셋 과 실험 데이터셋을 사용하여 제안된 기법을 실험하였다. 실험 결과는 이 논문에서 제안된 기법이 기존의 사각형 객체에 대한 선택율 추정기법보다 우수하다는 것을 입증하

였다. 특히, ②를 기반으로 하는 히스토그램의 경우 기존의 MinSkew, 웨이블릿 히스토그램보다 생성시간, 추정시간, 정확도가 높게 평가 되었으며, ①을 고려하는 히스토그램과는 생성 시간(약 0.26초)이나 추정시간(0.0004초)에서 약간의 오버헤드는 있지만, 선택율 면에서는 약 4%정도 좋은 성능을 나타내었다.

향후의 연구에서는 선택율의 정확성을 높일 수 있는 다른 확률모델을 고려하여 성능을 향상시킬 것이며, 크기가 증가하는 히스토그램을 효율적으로 압축할 수 있는 기법에 관한 연구도 수행할 것이다.

참 고 문 헌

[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), pp.299-310, Nov., 1995.

[2] Alberto Belussi, Christos Faloutsos, "*Self-Spatial Join Selectivity Estimation Using Fractal Concepts,*" In Proc. ACM Symp. on Transactions on Information Systems, Vol. 16, No.2, pp.161-201, April, 1998.

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

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

[5] Vitter, Wang, "*Approximate Computation of Multidimensional Aggregates of Sparse Data using Wavelets,*" In Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.193-204, 1999.

[6] C. Faloutsos, B. Seeger, A. Traina, and Caetano Traina, "*Spatial Join Selectivity Using Power Laws,*" In Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.177-188, 2000.

[7] A. Aboulnaga, J. Naughton, "*Accurate estimation of the cost of spatial selections,*" In Proceedings of the IEEE International Conference on Data Engineering(ICDE), pp.123-134, 2000.

[8] Yossi Matias, Jeffrey Scott Vitter, Min Wang, "*Dynamic Maintenance of Wavelet-Based Histograms,*" The VLDB, pp.101-110, Journal, 2000.

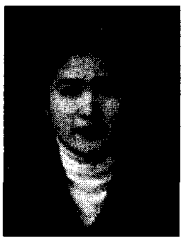
[9] Jin, N. An, A. Sivasubramaniam, "*Analyzing Range Queries on Spatial Data,*" In Proceedings of the IEEE International Conference on Data Engineering(ICDE), pp. 525-534, 2000.

[10] L. Getoor, B. Taskar, D. Koller, "*Selectivity estimationusing probabilistic models,*" In Proc. ACM SIGMOD Int. Conf. on Management of Data, pp.461-473, 2001.

[11] Min Wang, Jeffrey Scott Vitter, Lipyeow Lim, Sriram Padmanabhan, "*Wavelet-based cost Estimation for Spatial Queries,*" In Proc. Int. Symp. on Spatial and Temporal Databases, pp.175-196, 2001.

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

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



지 정 희

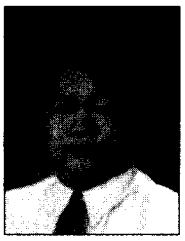
e-mail : jhchi@dblab.chungbuk.ac.kr

1999년 충주대학교 전자계산학과

2001년 충주대학교 대학원 전자계산학 석사

2003년 충북대학교 대학원 전자계산학 박사수료

관심분야 : 시공간 데이터베이스, Temporal GIS, 시공간 질의 최적화, 시공간 색인기법, 이동객체 관리 기법



김 상 호

e-mail : kimsh@dblab.chungbuk.ac.kr

1997년 충북대학교 컴퓨터과학과

1999년 충북대학교 대학원 전자계산학 석사

2004년 충북대학교 대학원 전자계산학 박사

2004년~현재 충북대학교 연구원

관심분야 : 시공간 데이터베이스, Web Visualization, Component GIS, 이동객체 관리기법

류 근 호

e-mail : khryu@dblab.chungbuk.ac.kr

1976년 숭실대학교 전자계산학과

1980년 연세대학교 공학대학원 전자계산학 석사

1988년 연세대학교 대학원 전자계산학 박사

1976년~1986년 육군군수지원사전산실(ROTC 장교), 한국전자통신연구소(연구원), 한국방송통신대 전산학과(조교수) 근무

1989년~1991년 Univ. of Arizona 연구원(TempIS Project)

1986년~현재 충북대학교 전기전자 및 컴퓨터공학부 교수

관심분야 : 시간 데이터베이스, 시공간 데이터베이스, Temporal GIS, 객체 및 지식베이스 시스템, 지식기반 정보검색시스템, 데이터마이닝, 데이터베이스 보안 및 Bio-Informatics