제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

영역객체의 공간 범위질의에 관한 선택률 추정기법 분석

정재혁*, 이진열, 지정희, 김상호, 류근호

충북대학교 데이터베이스 연구실
e-mail : {roknavy7, jinylee, jhchi, shkim, khryu}@dblab.chungbuk.ac.kr

Analysis of Selectivity Estimation Techniques for Spatial Range Query of Region Objects

Jae Hyuck Jeong*, Jin Yul Lee, Jeong Hee Chi, Sang Ho Kim, Keun Ho Ryu
Dept of Computer Science, Chungbuk National University

요 약

최근 공간 데이터베이스의 선택률 추정 문제에 대한 관심이 증가하면서, 데이터분포의 편중, 중복 계산, 메모리 공간 최소화등의 문제를 고려한 다양한 기법들이 제안되고 있다. 그러나 이들 기법들 간의 성능 분석을 통한 비교평가는 이루어지지 않고 있다. 따라서, 이 논문에서는 공간 영역객체의 범위질의에 관한 선택률 추정 기법인 Min-Skew, 웨이블릿, 누적밀도, 오일러 히스토그램을 비교 분석한다. 즉, 실제 데이터셋을 기반으로 여러 형태의 질의에 대한 성능 비교를 통해 각 기법들을 비교 평가한다. 이 연구 결과는 새로운 기법 도출에 이용될 수 있다.

1.서 론

최근 공간 데이터베이스에서 선택률 추정에 대한 관심이 증대되고 있다. 질의 결과를 추정하기 위한 대표적인 기법으로 샘플링, 파라메트릭, 히스토그램 기법 등이 있다. 이러한 기법들중 히스토그램은 적은 공간이 요구되고, 데이터 분포를 고려하지 않는 특징을 가짐으로써, 상용 데이터베이스에서 널리 사용되고 있다[1]. 하지만 공간 객체는 각기 다른 모양과 크기를 갖는 특성을 가지므로 기존 선택률 추정 기법을 공간 객체에 적용시 다음과 같은 문제점이 발생한다.

첫 번째, 기존의 히스토그램은 데이터의 분포가 모두 균일하다는 가정을 기반으로 한다. 하지만 이러한 가정은 실제 공간데이터가 항상 균일하게 분포한다고 보장할 수 없기 때문에 적용되기 어렵다. 만약 질의가 편중된 데이터 분포를 가진 버켓에 교차한다면, 질의에 대한 선택률은 큰 오차를 가져온다.

두 번째, 한정된 메모리 공간 때문에 공간이 커짐에 따라 필요한 공간 데이터의 요약 정보를 총 크기에 비해 적은 정보로 선택률을 추정함으로써, 신뢰성이 부족한 선택률을 가진다.

세 번째, 큰 공간 영역을 가지는 객체가 여러 그리드셀에 중복 계산됨으로써, 선택률에 오차를 가져올 수 있다.

이러한 히스토그램 기법의 문제점들을 해결하기 위한 다양한 기법들이 제안 되었다[1,2,5,6]. 이 논문에서는 선택률 추정시 공간 영역객체의 범위질의에 관한 연구로 제한을 두었다. 우선 첫 번째 문제점의 해결을 위해 공간 분포를 고려하여 공간 영역을 분할하는 Min-Skew 알고리즘이 제안되었고[1], 두 번째 문제점을 해결하기 위해 많은 양의 데이터를 압

*본 연구는 대학 IT연구센터 육성.지원사업의 연구 결과로 수행되었음.

축하여 데이터가 차지하는 메모리 공간을 감소시키는 웨이블릿 기반 히스토그램이 제안 되었다[6]. 마지막으로 영역객체의 중복문제를 해결하기 위해 누적밀도 히스토그램과 오일러 히스토그램이 제안되었다[2,5].

지금까지 공간 영역객체의 선택률 추정에 관한 다양한 기법들이 제안 되었지만, 이들 기법들에 관한 비교평가는 수행되지 않았다. 따라서, 이 논문에서는 공간 영역객체의 범위질의에 대한 선택률 추정 기법들을 분석 및 구현하여, 실제 데이터셋을 기반으로 여러 형태의 질의에 대한 성능 비교를 통해 각 기법들을 비교 평가한다. 이 논문의 목적은 각 히스토그램 기법의 특징과 문제점들을 실험을 통해 비교 분석함으로써, 새로운 히스토그램 기법 연구의 기반을 제공하는 것이다.

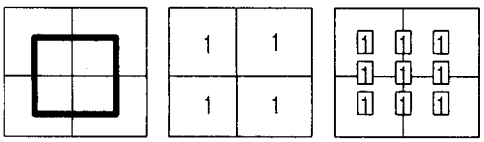
2장에서는 공간 선택률 추정을 위한 여러 가지 기법들에 대해 소개하고, 3장에서는 2장에서 소개되었던 각 히스토그램 기법들을 실험 평가하기 위해 평가 기준을 정의한다. 그리고 4장에서는 평가 결과를 통해 각 기법을 분석하고, 마지막으로 5장에서 결론을 맺는다.

2.공간 선택률 추정 기법

공간 선택률 추정이란 주어진 공간 질의와 교차하는 공간 객체의 수를 추정하는 것이다. 공간 데이터베이스에서 질의 최적화를 수행하기 위해 질의에 대한 모든 객체를 계산하는 것은 높은 비용을 초래한다. 따라서, 전체 데이터에 대한 요약정보를 구성하고, 이 요약정보를 기반으로 질의에 대한 추정된 선택률을 기반으로 최적화를 수행한다.

2.1.오일러 히스토그램

Beigel과 Tanin은 그래프 이론의 오일러 공식을 기반으로 공간 질의의 선택률 추정 알고리즘을 제안하였다[5].



a. 데이터셋 b. 전통적인 히스토그램 c. 오일러 히스토그램

(그림 1) 오일러 히스토그램의 예

그림1-a와 같은 데이터셋이 주어졌을 때, 전통적인 히스토그램은 그림1-b와 같이 중복 계산되는 문제가 발생한다. 오일러 히스토그램은 그림1-c와 같이 정점, 선분, 셀을 각각 계산함으로써, 이 문제를 해결한다.

질의윈도우 S가 주어지고, S의 선택률은 다음과 같이 계산된다.

$$Selectivit\ y(S) = \sum_{0 \le k \le d} (-1)^k F_k(S)$$

2.2.누적밀도 히스토그램

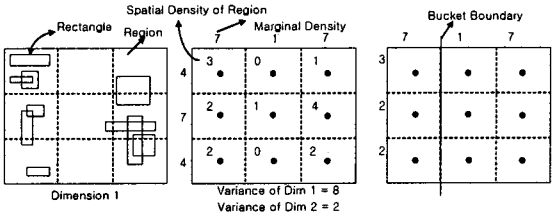
누적밀도 히스토그램은 공간 객체를 최소경계사각형(MBR)로 구성하여, 각 MBR의 좌하단 점, 우하단 점, 좌상단 점, 우상단 점과 교차하는 객체의 수를 4개의 서브히스토그램에 각각 저장하고, 유지하면서 누적 히스토그램을 생성하여 중복 카운트를 해결 하였다[2].

누적밀도 히스토그램은 질의에 대해 매우 빠른 처리와 높은 정확성을 보이지만 누적과 4개의 값을 유지해야하는데 필요한 메모리 비용이 높다는 단점을 가진다.

2.3.Min-Skew 히스토그램

Min-Skew 히스토그램은 공간 편중도를 최소화하기 위한 히스토그램이다[1]. 각 차원의 데이터 분포에 따라 공간 편중도가 큰 영역에 많은 버켓을 할당함으로써, 전체적인 편중도를 최소화한다. 그리고 분할을 결정하기 위해 그리디(Greedy) 알고리즘을 적용하여 계산 복잡도를 최소화 하였으며, 데이터를 처리하기 위해 전체 데이터를 메인 메모리에 유지하는 것을 요구하지 않는다.

그림2는 Min-Skew 처리과정을 보여준다.



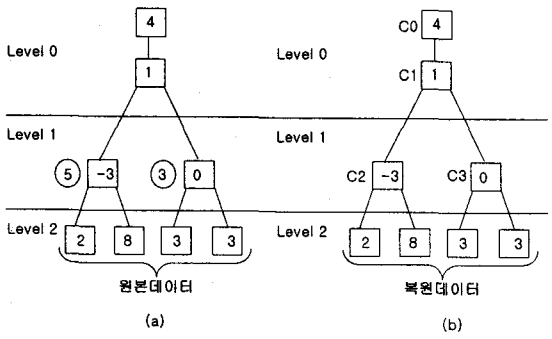
(그림 2) Min-Skew 히스토그램의 처리과정

2.4.웨이블릿

웨이블릿은 주로 이미지처리, 신호처리, 함수의 계층적 분석을 위한 수학적 도구로 사용되어 왔다[6]. 최근 이 기법을 데이터베이스의 근사 질의 처리와 선택률 추정기법에 적용하려는 많은 연구가 진행되고 있다.

<표 1> Haar 웨이블릿 예제

Resolution Level	Average	Detail Coefficient
0	[4]	[1]
1	[5 3]	[-3 0]
2	[2 8 3 3]	



(그림 3) 데이터 압축 및 복원

[6]의 Haar 웨이블릿은 표1을 그림 3과 같이 나타낼 수 있다. 원본데이터 {2,8,3,3}을 가지고, 두 쌍 {2,8}과 {3,3}을 각 평균값과 계수 값으로 유지한다. 두 쌍의 평균값과 계수값은 다음 공식으로 계산 된다 :

$$Average = \frac{a+b}{2}, \quad Difference = \frac{a-b}{2}$$

예1) 그림 3-a의 원본데이터 {2,8}은 평균값 5와 계수값 -3이 유지되고, 원본데이터 {3,3}은 각각 3과 0이 유지된다. 다음으로 평균값 5와 3을 가지고 평균값을 상위 루트 노드, 계수값을 하위 루트 누드에 유지한다. 여기서 계수값이 '0'이거나,'0'에 가까운 값은 데이터 복원에 영향을 주지 않으므로 저장하지 않는다. 따라서 이 과정을 통해 데이터 압축효과를 얻을 수 있으며, 이때 계수값을 '0'으로 만드는 과정을 웨이블릿 압축이라 한다.

예2) 그림 3-b은 압축된 데이터를 복원하는 예를 보여준다. 각 레벨에 유지되는 계수값을 통해 원본데이터를 복원 한다. 계수값C0{4}와C1{1}값의 합은 C2의 평균값을 의미하고 차는 C3의 평균값을 의미한다. C2의 평균값과 계수값의 합은 {2}, 차는{8}로 복원하게 된다.

3.실험 평가

이 논문의 실험은 펜티엄 III, 866MHz, Windows 2000 서버 환경에서 C++ 언어를 사용하여 Min-Skew, 누적밀도, 웨이블릿, 오일러 히스토그램 알고리즘을 구현하였다.

3.1.데이터셋 및 질의 윈도우

데이터셋은 11,000개의 서울시 중구 빌딩의 폴리곤 데이터셋을 사용하였고, 질의 윈도우는 5%에서 20% 사이에 변화를 주어 평가한다.

3.2.평가 기준

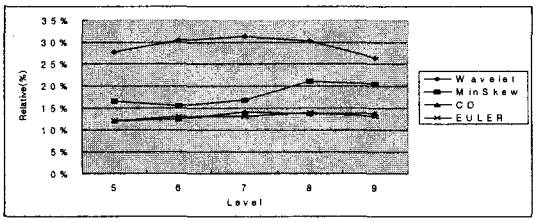
각 기법들의 효율성을 평가하기 위해, 다음과 같은 항목의 성능평가를 수행한다.

- ■상대적 평균 에러 비율
- ■질의 크기에 따른 에러 비율
- ■선택률 추정 시간
- ■히스토그램 생성 시간

4.실험 결과

4.1.상대적 평균 에러비율

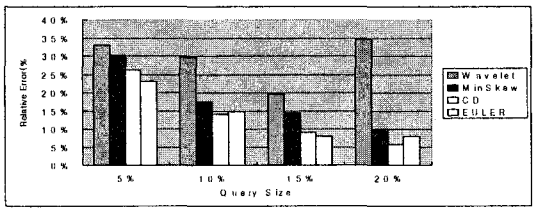
그림 4은 그리드레벨에 따른 각 기법의 상대적 평균에러 비율을 보여준다. 그래프의 x축은 히스토그램의 그리드 레벨을 나타내며, y축은 질의 윈도우로부터 계산된 상대적 에러비율을 의미한다. 실험결과 오일러 히스토그램과 누적밀도 히스토그램은 그리드 레벨에 따른 질의 윈도우로부터의 상대적 에러 비율이 거의 비슷하며, 특히 웨이블릿은 다른 히스토그램에 비해 에러 비율이 높은 것으로 나타났다.



(그림 4) 상대적 평균 에러 비율

4.2.질의 크기에 따른 에러 비율

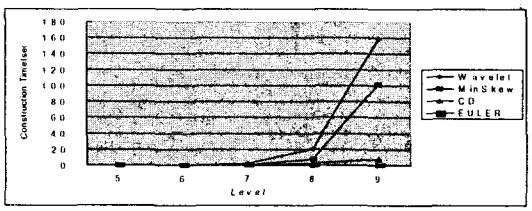
그림 5은 질의 크기에 따른 상대적 에러 비율의 보여준다. 대부분의 히스토그램이 질의 윈도우가 커짐에 따라 에러 비율이 확연히 낮아지는 것을 알 수 있었지만, 웨이블릿은 많은 변화가 일어나지 않는 결과를 보여준다.



(그림 5) 질의 크기 따른 평균 에러 비율

4.3.생성 시간 및 선택률 추정시간

그림 6에서 보여지는 것과 같이 레벨 7까지는 완만한 경사를 이루나 그 이상의 레벨에서는 급격히 생성시간이 증가하는 것을 알 수 있다. 표2에서는 질의 크기에 따른 선택률 추정시간을 측정하였다. 대부분의 히스토그램은 거의 비슷한 추정시간을 보인다. 웨이블릿은 1차원에 적합하기 때문에 2차원 형태로 공간 영역을 맵핑하는데 소요되는 비용으로 인해 다른 히스토그램에 비해 좀 더 느린 선택률 추정시간을 보였다.



(그림 6) 그리드 레벨에 따른 생성시간 측정

<표 2> 선택률 추정 시간 (단위 / 초)

구분/질의크기	5%	10%	15%	20%
Wavelet	0.0675	0.2297	0.4562	0.7087
MinSkew	0.0009	0.0006	0.0006	0.0006
CD	0.0000	0.0000	0.0000	0.0000
EULER	0.0072	0.0018	0.0020	0.0023

4.4.실험 결과 분석 및 평가

실험결과 다음과 같은 분석을 도출해 낼 수 있었다. Min-Skew 히스토그램은 공간 편중도를 최소화하기 위한 히스토그램이다. 그러나 객체의 삽입과 갱신에 의한 변화에 따라 공간 편중도가 변화 할 때마다 전체 공간을 다시 분할하는 단점으로 동적 구조에 부적합하며, 누적밀도 히스토그램이나 오일러 히스토그램에 비해 성능이 떨어지는 것을 확인할 수 있었다. 누적밀도 히스토그램과 오일러 히스토그램은 정확하고 빠른 수행이 기대되지만, 큰 메모리 공간이 요구되는 문제점을 가지고 있다. 또한 웨이블릿은 실험결과 다른 히스토그램에 비해 평균에러 비율과 수행 시간 측면에서 현저하게 좋지 않은 성능을 보인다. 왜냐하면, 웨이블릿은 1차원 데이터에 적합한 히스토그램이기 때문에 공간데이터에 적용하기 위해서는 공간 탐색 기법을 적용하여 2차원 형태로 공간 영역을 맵핑하는데 많은 시간과 비용이 소비되는 것으로 분석된다. 또한 웨이블릿은 중복 카운트 문제를 가지고 있기 때문에 정확한 추정이 어렵다. 우리의 연구 결과 각 히스토그램 기법은 서로 상호 보완적인 관계를 가지는 것을 발견 했고, 이러한 관계를 통해 둘 또는 그 이상의 기법을 결합하여 새로운 기법을 도출할 수 있을 것이라 분석된다.

5.결론 및 향후 연구

이 논문에서는 기존 히스토그램이 가지는 문제점을 해결하기 위해 제안된 각 히스토그램 기법들에 대해 알아보고, 각 기법을 구현하여 실험 평가를 통해 각 기법이 가지는 문제점과 특징을 비교 분석하였다. 이 연구 결과를 분석함으로써 각 기법의 특징과 문제점을 파악할 수 있었고, 이 파악된 결과를 이용하여 새로운 기법을 도출할 수 있는 토대를 마련하였다.

향후 연구로 누적밀도 히스토그램, 오일러 히스토그램, Min-Skew, 웨이블릿이 가지는 장점들은 확장하고, 문제점은 보완하여 보다 유연한 새로운 히스토그램 기법을 제안하는 연구가 진행될 것이다.

[참고문헌]

[1] 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.

[2] 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

[3] Nikos Mamoulis, Dimitris Papadias, "Selectivity estimation of complex spatial queries", In Proc. Int. Symp. on Spatial and Temporal Databases, 2001, pp. 156-174

[4] 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

[5] C. Sun, D. Agrawal, A. EI Abba야, "Selectivity for Spatial join with geometric selection", Proc. of EDBT, 2002, pp,609-626

[6] 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.