이동객체의 현재 질의에 대한 선택율 추정 기법

A Selectivity Estimation Technique for Current Query of Moving Object

최병갑, 지정희, 류근호
충북대학교 데이터베이스/바이오인포매틱스 연구실
choipower@paran.com, {jhchi, khryu}@dblab.cbu.ac.kr

Byung Kab Choi, Jeong Hee Chi, Keun Ho Ryu
Database/Bioinformatics Laboratory of Chungbuk National University

요 약 선택율 추정은 질의 최적화를 위한 기법중의 하나이다. 이동객체에 대한 기존 선택율 추정 기법은 시간에 따른 빈번한 이동객체의 위치 변화를 요약 정보에 반영하지 못함으로써 선택율 추정시 많은 에러를 발생시키고 있다. 따라서 이 논문에서는 이동객체의 질의에 대한 선택율 추정을 위한 색인 기반의 히스토그램 기법을 제안하였다. 또한 제안된 기법의 구현과 평가를 통해 제안된 기법의 성능을 분석하였다. 이 논문에서 제안된 기법은 차량 추적 시스템, 위치 기반 서비스, 응급 구조 서비스, 그리고 텔레매틱스 서비스 등과 같은 연속적으로 위치를 변경하는 이동객체의 정보를 실시간으로 관리하고 검색하는 응용분야에 활용 가능할 것이다.

1. 서 론

최근 무선 컴퓨팅 기술의 발달과 이동객체(moving object)의 위치를 실시간으로 추적할 수 있는 GPS(global positioning system) 기술의 발달로 인하여 물류 및 수송 관리, 디지털 전장, 항공 교통 통제, 위치 기반 서비스(LBS : location based service) 등과 같은 응용 서비스들이 개발되고 있다.

이동객체의 위치 정보를 관리하는 모바일(mobile) 데이터베이스 관리 시스템은 이러한 서비스의 효과적인 지원을 수행하기 위해 사용자가 요청한 다양한 질의를 빠른 시간 내에 처리할 수 있어야 한다. 아울러 시스템은 시간의 흐름에 따라 연속적으로 변화하며, 매우 빈번하게 갱신되는 특성을 갖는 이동객체의 위치 정보를 신속하게 관리하여야 한다. 이 논문에서는 이동객체의 현재 위치를 기반으로 수행되는 현재 질의를 신속하게 처리하기 위하여 쿼드 트리 기반의 히스토그램, *QTH*(quad tree based histogram)을 이용한 선택율 추정(selectivity estimation) 기법을 제안한다.

선택율이란 전체 데이터 셋 중에서 질의를 만족하는 데이터의 수 혹은 비율을 의미한다[1]. 최근까지 많은 공간 선택율 추정 기법들이 제안되어 왔다. 그러나 기존 공간 질의 선택율 추정과 이동객체 질의 선택율 추정 기법은 각 기법에서 다루는 데이터 셋의 특성이 다르다. 즉 공간 선택율 추정에서의 공간객체는 객체의 위치 정보가 정적이지만, 이동객체는

이 연구는 정보통신부 대학 IT연구센터 육성 및 지원사업의 연구비 지원으로 수행되었음.

시간에 따라 위치 정보가 연속적으로 변하는 특성을 갖는다. 따라서 이동객체의 정보를 요약하기 위해 사용된 기존 공간 히스토그램[2,3,4,5]은 이동객체의 현재 위치 정보에 대한 범위 질의, 즉 현재 질의에 대한 선택율 추정치를 제공해 줄 수 있지만, 시간에 따른 객체의 위치 변화를 반영하기 어렵다는 문제점을 갖는다.

최근 여러 기법들이 시간에 따라 정보를 변화시키는 이동객체의 특성을 요약 정보에 반영하기 위해 제안되고 있다[6,7,8,9]. 기존 기법들은 공간 분할 알고리즘을 기반으로 히스토그램을 생성하거나, 차원변환을 이용하여 히스토그램을 생성하고, 공간 색인을 기반으로 히스토그램을 생성하고 있다. 그러나 이들 기법의 버켓내 균일성은 초기 데이터 분포의 공간 분할 정책에 따라 생성되므로, 이동객체의 위치 정보가 변함에 따라서 쉽게 깨어질 수 있다. 이러한 문제점은 히스토그램을 자주 재생성함으로써 해결할 수 있지만, 빈번한 히스토그램 재생성으로 인한 오버헤드를 초래하게 된다. 그러므로 시간에 따른 이동객체의 갱신정보를 반영하여 이동객체에 대한 현재 질의에 대한 정확한 선택율 추정 결과를 제공해 줄 수 있는 기법이 필요하다.

2. 관련 연구

이동객체의 현재 위치 정보는 기존 정적인 공간 객체에 대한 위치 정보와 동일한 개념으로 다루어 질 수 있다. 여러 연구들이 공간 객체의 범위 질의에 대한 선택율을 추정하기 위해 수행되어 왔다. 이들 기법들은 공간 분할 정책 및 버켓에 유지되는 정보에 따라 각기 다른 성능을 나타낸다.

Acharya[3]는 특히 편중된 공간 분포를 갖는 데이터 셋에 적용하기 위해 MinSkew 히스토그램 기법을 제안하고 있다. 이 기법의 목적은 버켓 내의 공간 분포를 최대한 균등하게 함으로써 버켓 내의 균등성 가정으로 인한 선택율 에러를 줄이는 것이다. 그러나 이 기법은 공간 분할시 축 분할을 사용하기 때문에 불필요한 버켓들을 생성하는 문제점을 갖는다. 또한 공간 영역 객체는 버켓내에 포함되지 않고, 여러 버켓에 걸쳐지게 되는 경우 버켓 내의 객체 수를 다중 계산하는 문제를 발생하게 된다. 이러한 문제를 해결하기 위해 Jin[4]은 CD(cumulative density) 히스토그램을 제안하였다. 이 기법은 영역 객체의 최소 경계 사각형(MBR : minimum boundary rectangle)의 네 모서리 점을 하부 누적 히스토그램에 각각 저장한다. 따라서 이 기법은 한 객체에 대해 네 번 저장하여야 하는 부담을 가진다.

이들 공간 히스토그램 기법들은 객체의 현재 공간 위치 정보를 기반으로 공간 분할 작업을 수행한다. 그러므로 이들 기법이 시간에 따라 변하는 이동객체의 위치 정보를 반영하기 위해서는 히스토그램 생성에 이용된 공간 분할 정책이 객체의 갱신 정보를 반영할 수 있도록 유연한 구조를 가져야 한다.

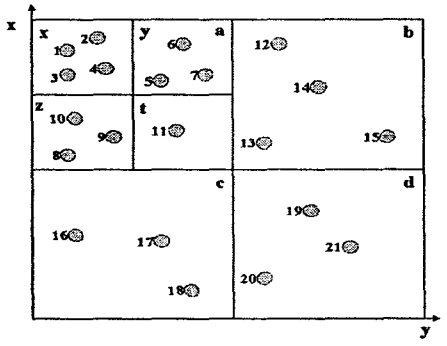
이러한 히스토그램 갱신 문제는 색인을 이용한 히스토그램 기법으로 해결 될 수 있다. 황규영[2]는 MLGF(multi level grid file)라는 공간 색인 구조에 카운트(count) 필드를 추가하여 선택율을 추정하였다. MLGF는 동적인 계층 평준화가 이루어지는 다차원 파일 구조이기에 비균일 데이터 분포에 대해서도 좋은 성능을 보인다. 그러나 이 기법은 선택율 추정을 위하여 많은 양의 공간 색인 구조를 읽어야 하기 때문에 많은 디스크 접근이 필요하여 추정 시간이 증가한다는 단점이 있다. 또한 선종복[5]은 쿼드 트리를 이용한 통계 데이터 관리 기법을 제안하고 있다. 이들 기법은 초기 데이터 분포를 반영한 통계 데이터 저장 구조를 만들고, 이 데이터 구조를 기반으로 쿼드 트리 분할 알고리즘을 적용하여 통계 데이터를 생성하고 있다. 그러나 이 기법은 데이터의 갱신이 일어나 분할과 병합이 발생할 경우, 공간

데이터를 다시 읽어 들여 통계데이터를 다시 생성해야 하는 문제점이 있다.

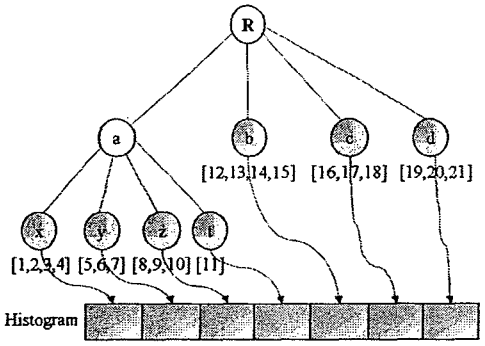
따라서, 이 논문에서는 이러한 기존 기법의 문제점을 해결하기 위하여 이동객체를 위한 쿼드 트리 기반의 선택율 추정 기법을 제안한다.

3. 제안된 기법의 구조

이 논문에서 제안하는 쿼드 트리 기반의 히스토그램, QTH의 구조는 그림 1과 같다.



(a) 데이터 셋



(b) 히스토그램

그림 1. QTH의 구조

QTH는 쿼드 트리와 해쉬 테이블을 기반으로 구성된다. 쿼드 트리[10]는 노드 간의 겹침이 존재하지 않는 색인 기법으로, 공간상에 위치한 객체의 수를 기반으로 비균등 공간 분할 방식을 통해 노드를 생성하므로 그리드 기반의 색인 기법에 비해 데이터 분포를 잘 반영할 수 있다. 또한 이 기법은 단순한 분할 및 병합 정책을 지원하므로 기존의 객체 기반 공간 분할 정책을 지원하는 색인 기법인 R-Tree 계열의 TPR-Tree, TPR*-Tree, HTPR-Tree 등의 색인 기법에 비해 간단히 구현되고 유지될 수 있다. QTH에서의 쿼드 트리는 히스토그램의 버켓 생성을 위한 분할 알고리즘을 제공하며, 또한 객체의 갱신 정보를 히스토그램에 반영하는 갱신 알고리즘을 제공한다. 해쉬 테이블은 쿼드 트리의 노드에 해당하는 버켓으로 구성된 히스토그램을 유지한다. 쿼드 트리를 기반으로 하는 히스토그램은 색인을 생성할 때 히스토그램도 동시에 생성할 수 있으므로, 히스토그램을 구성하기 위해 데이터베이스 전체를 다시 스캔해야 할 필요가 없다.

QTH는 쿼드 트리 생성, 버켓 생성, 요약 정보 저장, 히스토그램 조정 단계를 거쳐 생성된다. 각 절차에 관한 내용을 정리하면 다음과 같다.

먼저, 이동객체 데이터를 기반으로 쿼드 트리를 생성한다. 히스토그램을 생성하기 위해, 전체 데이터 셋을 기반으로 쿼드 트리 색인을 구성하며, 객체의 공간 차원을 기반으로 분할 작업을 수행한다. 비균등 분할 방법은 객체의 삽입 및 삭제가 빈번한 경우 각 노드의 영역을 재구성하는 것이 균등분할 방법보다는 복잡하다. 우선 객체가 삭제되어 분할된 셀의 카운트의 합이 제한 개수보다 작아지는 경우에는 셀을 병합해서 1개의 카운트 필드를 유지하게 된다. 또한, 한 셀에 삽입이 계속 발생하는 경우, 셀에 속하는 객체를 다시 읽어 여러 개의 셀로 나누어 주는 동적 분할 및 병합을 수행해야 하는 단점이 있다. 그러나 전체 데이터가 아닌 해당 영역의 객체만 다시 읽어주면 되기 때문에 처리 시간이 길지 않다. 또한 이러한 작업은 질의 시간이 아닌 객체의 변경시간에 수행되므로, 질의 선택율 추정시 영향을 주지 않는다.

QTH를 위한 쿼드 트리의 노드 구조는 <*numberOfObjects*, *nodeArea*, *cPointer*, *level*, *flag*, *objectList*>로 구성된다. 여기서 *numberOfObjects*는 현재 노드에 포함되어 있는 이동객체의 수를 의미하고, nodeArea는 노드의 영역 정보를 의미한다. 그리고 cPointer는 자식 노드를 가리키는 포인터이며, flag는 단말 또는 비단말 노드를 구별하기 위한 속성이며, objectList는 노드에 포함되어 있는 객체 리스트를 의미한다.

다음 단계에서는 쿼드 트리의 노드를 생성 후 히스토그램의 버켓을 생성한다. 히스토그램의 버켓은 쿼드 트리의 단말노드 식별자를 해쉬 키로 이용하여 단말 노드에 대응하는 버켓을 유지함으로써 생성된다. 이 기법은 해쉬 테이블을 기반으로 히스토그램을 유지함으로써, 편중된 데이터일 경우 트리 높이에 따른 검색 오버헤드 문제를 해결할 수 있다. 또한 객체의 갱신이 발생할 경우, 색인의 단말노드에 반영된 객체의 갱신 정보를 간단히 해쉬 키에 대응하는 버켓에 반영할 수 있다.

버켓 생성 후, 각 버켓에는 <*nid*, *MBR*, *numberOfObjects*> 정보를 저장한다. 여기서 nid는 버켓 식별자를 나타내며, MBR은 버켓 내의 공간 경계 영역을 나타내며, numberOfObjects는 버켓 내의 포함된 객체의 수를 의미한다.

쿼드 트리는 노드에서 수용할 수 있는 객체의 수를 초과할 때 공간 분할을 수행함으로 해서, 객체가 존재하지 않는 빈 노드가 생성된다. 이러한 빈 노드에 버켓을 할당할 경우 불필요한 버켓들이 히스토그램을 구성하게 되므로, 이러한 빈 노드에 대한 버켓을 생성하지 않기 위해 객체의 수가 0인 노드의 경우 버켓을 생성시키지 않는다. 객체의 수가 0인 버켓을 생성하지 않으므로 해서 선택율 추정시 불필요한 연산을 수행하지 않도록 한다.

4. 실험 및 평가

이 논문에서 제안된 *QTH* 기법의 성능을 평가하기 위하여 MinSkew 기반 공간 히스토그램, *MH* 및 R*-Tree 기반의 히스토그램, *RH*와 다양한 실험 요인들을 변화시키면서 비교 평가 하였다.

히스토그램의 성능은 데이터 분포에 상당히 의존적이다. 이 논문에서는 데이터 셋의 분포를 변화하여 실험 데이터 셋을 생성하기 위해 GSTD[11]를 이용하였다. 실험 데이터 셋은 GSTD를 이용하여 편중된 분포를 갖는 객체 100,000개를 생성하여 이용하였다. 이때 편중된 데이터 셋의 편중 계수(skewness coefficient)는 0.8로 설정하여 생성하였으며, 생성된 데이터는 공간 범위 <(0,0), (10000, 10000)> 내에 존재한다.

히스토그램 기법의 목적은 정확한 선택율을 추정하는 것이다. 따라서 정확도는 히스토그램의 성능을 평가하기 위한 중요한 요소가 된다.

실험 결과인 선택율 추정치의 정확도를 평가하기 위해서 식(1)과 같이 정의된 평균 상대 에러 E_r를 사용하였다.

$$E_r = \frac{\left(\sum_{q_i=1}^{q_{i=n}} |S(q_i) - S'(q_i)|\right)}{\sum_{q_i=1}^{q_{i=n}} S(q_i)} \times 100 \qquad \text{식(1)}$$

식(1)에서 qi는 i번째 질의를 의미하며, 실험에서는 1,000개의 질의에 대한 평균 에러로 정확도를 평가하였다. S(qi)는 질의 qi에 대한 실제 선택율을 나타내며, S'(qi)는 히스토그램을 이용하여 추정한 선택율을 의미한다.

이 논문에서 제안된 기법의 성능을 평가하기 위하여, 질의 크기, 저장공간의 크기, 데이터의 갱신 비율에 따른 정확도를 평가하였다. 히스토그램의 저장 공간에 따른 정확도를 평가하기 위해 300, 600, 1000개의 버켓들로 구성된 3 종류의 히스토그램을 생성하였으며,

질의의 크기에 따른 성능을 평가하기 위해 GSTD를 이용하여 전체 공간 영역의 5%~20% 크기를 갖는 범위 질의 1,000개씩을 생성하여 실험하였다.

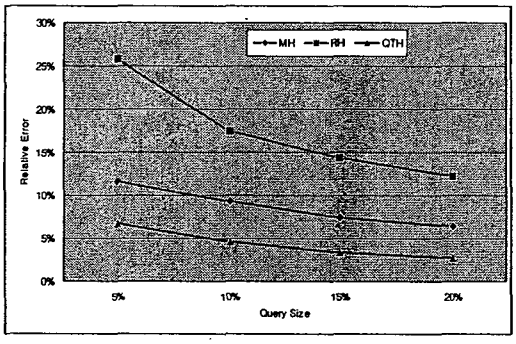


그림 2. 질의 크기에 따른 정확도

그림 2는 5%~20% 질의에 대한 평균 추정 결과이다. 실험결과는 모든 기법들이 질의의 크기가 증가할수록 선택율 추정의 정확도가 증가하는 것으로 나타났으며, 특히 이 논문에서 제안한 *QTH* 기법이 가장 높은 정확도를 보이고 있다. 이러한 이유는 *R*-Tree* 기반의 *RH* 기법의 경우 트리의 중간 노드에 버켓을 할당하여 히스토그램을 생성하게 되므로, 버켓 사이의 중복이 발생하므로, 선택율 추정시 교차하는 버켓에 대한 선택율을 중복 계산하게 되는 문제를 발생시킨다. 따라서 추정 결과는 실제 선택율과 많은 에러를 발생시키게 된다.

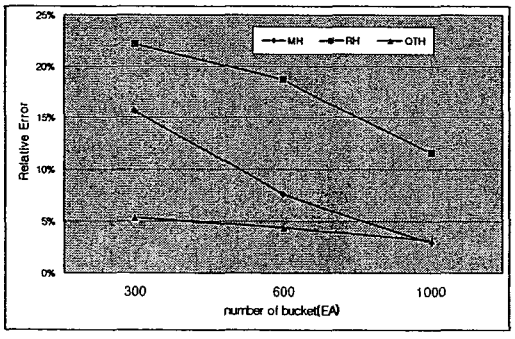


그림 3. 저장공간 크기에 따른 정확도

그림 3은 저장공간 크기에 따른 정확도 실험 결과를 보여주고 있다. 실험 결과는 모든 기법이 버켓의 수가 증가할수록 정확도가 증가하는 경향을 보이고 있다. 이는 버켓의 수가 증가할수록 데이터 분포를 잘 반영할 수 있기 때문이다. 또한 실험 결과는 이 논문에서 제안한 *QTH* 기법이 가장 좋은 성능을 가짐을 보여주고 있다. 이러한 이유는 *QTH* 기법이 분할된 공간 내의 객체가 없을 경우 버켓을 생성하지 않음으로서 다른 기법에 비해 데이터 분포를 반영할 수 있는 버켓들을 많이 확보할 수 있기 때문이다.

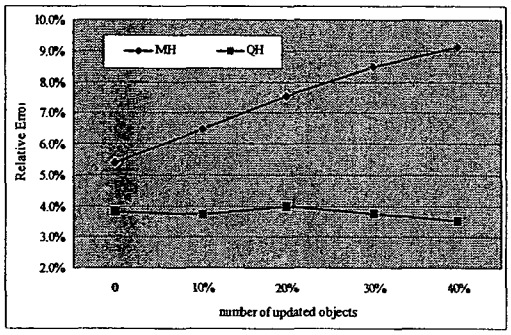


그림 4. 데이터 갱신에 따른 정확도

그림 4는 이동객체의 갱신에 따른 히스토그램 성능의 평가 결과를 보여주고 있다. 이 실험에서는 갱신에 따른 정확도를 평가하기 위해, 전체 데이터의 10~40%까지 갱신한 데이터를 기반으로 실험하였다. 실험 결과는 이 논문에서 제안한 *QTH* 기법의 경우 갱신에 따른 정확도에 큰 차이를 보이지 않은 반면, *MinSkew* 기반의 *MH* 기법은 갱신에 따라 정확도가 떨어짐을 보이고 있다. 이러한 이유는 *QTH* 기법은 쿼드 트리의 갱신 정책에 따라 히스토그램이 갱신되지만, *MH* 기법의 경우 갱신 이전에 데이터 셋을 기반으로 분할된 공간 영역이 갱신된 객체의 공간 분포를 반영하지 못하기 때문이다. 이 기법에서 갱신된 데이터의 공간 분포를 히스토그램에 반영하기 위해서는 히스토그램을 다시 재생성해야 한다.

5. 결론

선택율 추정은 질의 최적화를 위한 기법중의 하나이다. 이동객체의 질의에 대한 기존 선택율 추정 기법은 시간에 따른 이동객체의 위치 변화를 요약 정보에 반영하지 못하며, 또한 기존 공간 요약 정보를 확장하여 이용함으로써 선택율 추정시 많은 에러를 발생시키고 있다.

따라서 이 논문에서는 이동객체 질의 선택율 추정 기법을 개발하기 위하여 쿼드 트리 기반의 히스토그램 기법을 제안하였다. 제안된 기법은 쿼드 트리를 기반으로 생성되므로, 이동객체의 빈번한 위치 갱신 정보를 히스토그램에 쉽게 반영할 수 있으며, 따라서 히스토그램을 빈번히 재생성할 필요가 없다. 다양한 실험 결과는 이 논문에서 제안한 쿼드 트리 기반 히스토그램, *QTH* 기법이 우수한 성능을 가짐을 보였다.

이 논문에서 제안된 기법은 차량 추적 시스템, 위치 기반 서비스, 응급 구조 서비스, 그리고 텔레매틱스 서비스 등과 같은 연속적으로 위치를 변경하는 이동객체의 정보를 실시간으로 관리하고 검색하는 응용분야에 활용 가능하다.

참고문헌

[1] Y. E. Ioannidis, "Query Optimization", ACM Computing Surveys, Vol.28, No.1, pp.121-123, Mar. 1996.

[2] K. Y. Whang, S. W. Kim, and G. Wiederhold, "Dynamic Maintenance of Data Distribution for Selectivity Estimation," The VLDB Journal, Vol.3, No.1, pp.29-51, Jan. 1994.

[3] S. Acharya, V. Poosala, and S. Ramaswamy, "Selectivity Estimation in Spatial Databases", ACM SIGMOD, pp.13-24, Jun. 1999.

[4] J. Jin, N. An, and A. Sivasubramaniam, "Analyzing Range Queries on Spatial Data", ICDE, pp.525-534, Feb. 2000.

[5] 선종복, 김진덕, 김장수, 홍봉희, "공간질의 최적화를 위한 선택률 추정 방법," 정보과학회 논문지(D), 제25권 제7호, pp980-995, 1998년 7월.

[6] Y. J. Choi, and C. W. Chung, "Selectivity Estimation for Spatio-Temporal Queries to Moving Objects", ACM SIGMOD, pp.440-451, Jun. 2002.

[7] M. Hadjieleftheriou, G. Kollios, and V. Tsotras, "Performance Evaluation of Spatio-Temporal Selectivity Estimation Techniques", SSDB, pp.202-211, Jul. 2003.

[8] Y. Tao, J. Sun, and D. Papadias, "Selectivity Estimation for Predictive Spatio-Temporal Queries", ICDE, pp.417-428, Mar. 2003.

[9] H. G. Elmongui, M. F. Mokbel, and W. G. Aref, "Spatio-temporal Histogram", SSTD, pp.19-36, Aug. 2005.

[10] H.Samet, "The Design and Analysis of Spatial Data Structures," Addison Wesley Publishing Company, INC., 1994.

[11] Y. Theodoridis, J. R. O. Silva, and M. A. Nascimento, "On the Generation of Spatiotemporal Datasets", SSDB, Hong Kong, China, pp.147-164, Jul. 1999.