

이동객체를 위한 현재 질의 선택을 추정 기법

지정희*, 류근호**, 정두영***

A Selectivity Estimation Technique for Current Query of Moving Objects

Jeong-Hee Chi *, Keun-Ho Ryu **, Doo-Young Jeong ***

요 약

선택을 추정은 질의 최적화를 위한 기법중의 하나이다. 이동객체에 대한 기존 선택을 추정 기법은 시간에 따른 이동객체의 위치 변화를 요약 정보에 반영하지 못하며, 또한 기존 공간 요약 정보를 확장하여 이 용함으로써 선택을 추정시 많은 에러를 발생시키고 있다. 기존 기법들이 이동객체의 위치 정보 변화를 요약 정보에 반영하기 위해서는 요약 정보를 자주 재생성해야 하며, 그러므로 전체 데이터베이스를 자주 스캔해야 하는 문제점을 갖고 있다. 따라서 이 논문에서는 이동객체의 현재 질의에 대한 선택을 추정 기법을 개발하기 위하여 쿼드 트리 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.

Abstract

Selectivity estimation is one of the query optimization techniques. It is difficult for the previous selectivity estimation techniques for moving objects to apply the position change of moving objects to synopsis. Therefore, they result in much error when estimating selectivity for queries, because they are based on the extended spatial synopsis which does not consider the property of the moving objects. In order to reduce the estimation error, the existing techniques should often rebuild the synopsis. Consequently problem occurs, that is, the whole database should be read frequently. In this paper, we proposed a moving object histogram method based on quad tree to develop a selectivity estimation technique for moving object queries. We then analyzed the performance of the proposed method through the implementation and evaluation of the proposed method. Our method can be used in various location management systems such as vehicle location tracking systems, location based services, telematics services, emergency rescue service, etc in which the location information of moving objects changes over time.

▶ Keyword : Selectivity Estimation, Spatiotemporal Selectivity, Moving Objects, Histogram

• 제1저자 : 지정희

• 접수일 : 2006.01.07, 심사완료일 : 2006.02.24

* 충북대학교 전기전자컴퓨터공학부 박사, **,*** 충북대학교 전기전자컴퓨터공학부 교수

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

I. 서론

최근 무선 컴퓨팅 기술(wireless computing technology)의 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)를 이용한 선택을 추정(selectivity estimation) 기법을 제안한다.

선택율이란 전체 데이터 셋 중에서 질의를 만족하는 데이터의 수 혹은 비율을 의미한다. 최근 데이터베이스의 크기가 급격히 증가하면서 데이터베이스 시스템의 질의 최적화(query optimizer)나 질의 프로파일러(query profiler) 모듈들이 질의 결과에 대한 정확한 선택을 추정치를 요구하고 있다[1,2]. 질의 최적화기는 대규모 데이터베이스 질의에 대한 효율적인 실행 계획을 선택하기 위하여 전체 데이터베이스를 접근하는 비효율적인 방식 대신, 요약정보(synopsis)를 기반으로 수행된 선택을 추정 결과를 사용하여 최적의 실행 계획을 선택한다[3]. 그리고 질의 프로파일러는 질의가 실제로 수행되기 전 선택을 추정 결과를 사용하여 사용자에게 실행시간 정보를 제공한다[4]. 또한 데이터웨어하우스(data warehouse)에서는 질의에 대한 정확한 응답을 얻기 위해서 많은 시간과 공간을 필요로 하므로, 선택을 추정 기법을 이용하여 빠른 근사 질의 처리를 수행함으로써 OLAP(online analytical processing) 질의를 수행하고 있다[5].

대부분의 데이터베이스 관리 시스템들은 효과적인 질의 처리를 위하여 요약 정보를 사용한다. 요약 정보는 데이터베이스에 저장된 레코드들의 특정 속성값에 대한 데이터 분포를 함축적으로 표현한 것으로서 해당 속성값과 빈도의 쌍으로 이루어진다[6]. 요약 정보를 구성하기 위한 기법들로 히스토그램 기법, 샘플링 기법, 파라미터 기법이 연구되었다. 여러 기법 중에 히스토그램 기법은 매우 적은 공간을 사용하고, 사전에 데이터 분포를 알고 있지 않아도 되므로 많은 상용 데이터베이스 시스템에서 사용되고 있다.

그러나 기존 공간 질의 선택을 추정과 이동객체 질의 선택을 추정 기법은 각 기법에서 다루는 데이터 셋의 특성이 다르다. 즉 공간 선택을 추정에서의 공간객체는 객체의 위치 정보가 정적이지만, 이동객체는 시간에 따라 위치 정보가 연속적으로 변하는 특성을 갖는다. 따라서 이동객체의 정보를 요약하기 위해 사용된 기존 공간 히스토그램[7,8,9,10]은 이동객체의 현재 위치 정보에 대한 범위 질의, 즉 현재 질의에 대한 선택을 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[11,12,13,14]. 기존 기법들은 공간 분할 알고리즘을 기반으로 히스토그램을 생성하거나, 차원변환을 이용하여 히스토그램을 생성하고, 공간 색인을 기반으로 히스토그램을 생성하고 있다. 그러나 이들 기법의 버킷내 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 선택을 추정 결과를 제공해 줄 수 있는 기법이 필요하다.

II. 관련연구

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 버킷에 유지되는 정보에 따라 각기 다른 성능을 나타낸다.

Acharya[7]는 특히 편중된 공간 분포를 갖는 데이터 셋에 적용하기 위해 MinSkew 히스토그램 기법을 제안하고 있다. 이 기법의 목적은 버킷 내의 공간 분포를 최대한 균등하게 함으로써 버킷 내의 균등성 가정으로 인한 선택을 에러를 줄이는 것이다. 그러나 이 기법은 공간 분할시 축 분할을 사용하기 때문에 불필요한 버킷들을 생성하는 문제점을 갖는다. 또한 공간 영역 객체는 버킷내에 포함되지 않고, 여러 버킷에 걸쳐지게 되는 경우 버킷 내의 객체 수를 다중 계산하는 문제를 발생하게 된다. 이러한 문제를 해결하기 위해 Jin[10]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. 이 밖에 선택을 추정 위한 다른 기법으로는 차원 변환을 이용한 기법(8,9,15)이 존재한다. 특히 Wang[15]은 웨이블릿 히스토그램과 객체의 최소 경계 사각형의 각 변을 점으로 차원 변환하여 비교적 낮은 저장 공간을 유지하면서 선택을 에러가 적은 히스토그램을 제안하였다. 그러나 이 기법은 객체의 최소 경계 사각형 네 변에 대해 각각 히스토그램을 유지하는 단점이 있다.

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.

기존의 MinSkew 히스토그램 기법은 매우 정적인 분할 구조를 갖는 기법이며, 데이터의 균등성을 고려하여 분할된 공간 영역은 객체의 이동으로 쉽게 깨어질 수 있다. 따라서 이 기법은 객체의 갱신 정보를 요약 정보에 반영하기 위해서는 전체 데이터 셋을 읽어서 다시 공간 분할 작업을 수행해야 한다. 또한 누적 밀도 히스토그램과 같은 고정 그리드 셀로 분할하여 작업을 수행하는 기법은 그리드 셀의 분할 레벨에 따라 성능이 결정된다. 즉, 그리드 셀을 세밀히 분할하는 경우 정확한 선택을 추정할 수 있지만, 그만큼의 많은 저장 공간을 요구한다. 그리고 공간 변환 기법을 기반으로 하는 기법은 데이터 생성 및 선택을 추정시 데이터의 차원을 변환시키는데 많은 오버헤드를 초래하게 된다.

이러한 히스토그램 갱신 문제는 색인을 이용한 히스토그램 기법으로 해결 될 수 있다. 황규영[6]은 MLGF(multi

level grid file)라는 공간 색인 구조에 카운트(count) 필드를 추가하여 선택을 추정하였다. MLGF는 동적인 계층 평준화가 이루어지는 다차원 파일 구조이기에 비균일 데이터 분포에 대해서도 좋은 성능을 보인다. 그러나 이 기법은 선택을 추정을 위하여 많은 양의 공간 색인 구조를 읽어야 하기 때문에 많은 디스크 접근이 필요하여 추정 시간이 증가한다는 단점이 있다. 또한 선중복[16]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 선택을 추정 기법을 제안한다.

III. 쿼드 트리 기반 히스토그램 기법

3.1 이동객체의 현재 질의

이동객체는 시간에 따라 위치가 변한다. 현재 질의는 이러한 이동객체의 현재 위치 정보에 대한 범위 질의를 의미한다. 이동객체의 현재 위치 정보에 대한 “현재 영화관에서 반경 500m 이내에 위치한 고객을 검색하십시오” 같은 질의는 이동 단말기를 소지한 사용자에게 쿠폰 서비스를 제공하기 위해 자주 사용되는 질의 유형이다.

(그림 1)은 이동객체의 현재 질의 예를 보여주고 있다. 현재 시간 $vt=0$ 에서, 객체의 위치 정보가 (그림 1(a))와 같을 때, “현재 주어진 윈도우 영역 내에 위치한 차량을 검색하십시오”와 같이 객체의 정보가 저장된 시간 vt 와 질의 시간 qt 이 동일한 질의가 주어지면, (그림 1(b))와 같이 객체의 현재 위치 정보를 기반으로 객체 C 와 D가 질의를 만족한다고 응답한다.

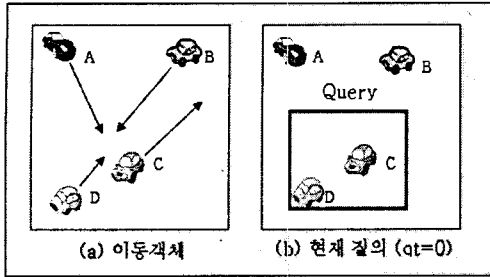


그림 1. 이동객체의 현재 질의
Figure 1. Current Query of Moving Objects

이동객체 현재 질의에 대한 선택율은 이동객체의 현재 위치 정보에 대한 범위 질의를 만족하는 객체의 수를 의미한다. (그림 2)에서와 같이 전체 공간 U 에 N 개의 객체가 균일하게 존재한다고 할 때, 현재 질의에 대한 선택율 S_Q 는 전체 공간 영역의 크기를 $Ared(U)$ 라 하고, 질의 Q_c 의 영역 크기를 $Ared(Q_c)$ 라 할 때, 두 영역 사이의 교차 비율이 질의를 만족하는 객체의 수라고 추정할 수 있다.

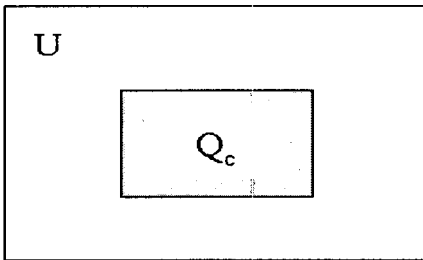


그림 2. 균일 데이터에 대한 현재질의
Figure 2. Current Query for Uniform data Distribution

그러나 실세계 객체의 대부분이 균일한 데이터 분포를 갖는다고 할 수 없으므로, 가능한 실세계 데이터들이 균일한 분포를 갖도록 (그림 3)과 같이 공간 분할 작업을 수행한다.

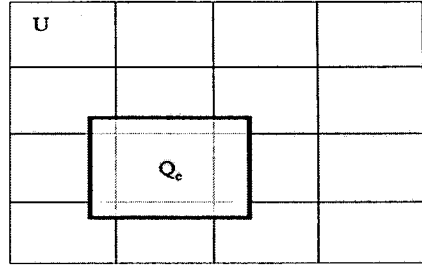


그림 3. 분할된 영역에서의 현재질의
Figure 3. Current Query in Partitioned Data Space

공간 분할 작업은 균등 분할과 비균등 분할 방식으로 구분될 수 있다. 균등 분할방식은 모든 버킷의 크기가 균일하도록 분할하는 방식으로, 구현 및 유지하기 쉽다는 장점을 갖는 반면, 데이터 분포에 상관없이 균일한 형태로 분할하게 되므로 데이터의 분포를 반영하지 못한다는 단점을 갖는다. 또한 비균등 분할 방식은 데이터의 분포를 반영한 요약 정보를 생성할 수 있는 반면, 균등 분할 방식에 비해 복잡한 작업을 수행해야 하는 단점을 갖는다.

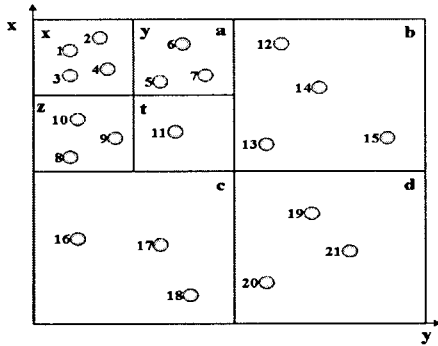
일반적으로 현재 질의에 대한 선택율 S_Q 는 식 (1)과 같이 균등 및 비균등 방식으로 분할된 각 영역과 질의와의 교차 비율에 의해 추정된다.

$$S_{Q_c} = \sum_{i=1}^{i=k} N_i * \frac{Ared(n_i \cap Q_c)}{Ared(n_i)} \dots\dots\dots (식 1)$$

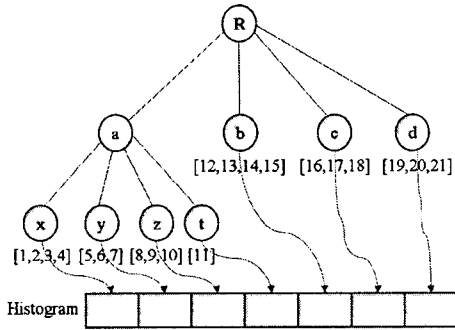
즉, 전체 버킷의 개수가 k 개이고, i 번째 버킷을 n_i 라 하고, 이 버킷에 속하는 객체의 수를 N_i 라 할 때, 현재 질의에 대한 선택율 S_Q 는 각 버킷과 질의와의 교차 비율만큼 객체가 존재한다고 추정한다. 다시 설명하면, 버킷의 영역 크기는 $Ared(n_i)$ 이고, 이 버킷과 질의 사이의 교차영역 크기는 $Ared(n_i \cap Q_c)$ 라 할 때, 교차 영역 내에 객체가 포함될 확률은 $Ared(n_i \cap Q_c) / Ared(n_i)$ 이 된다.

3.2 제안된 기법의 구조

이 논문에서 제안하는 쿼드 트리 기반의 히스토그램, QTH의 구조는 (그림 4)와 같다.



(a) 데이터 셋



(b) 히스토그램

그림 4. QTH의 구조
Figure 4. The Structure of QTH

*QTH*는 쿼드 트리와 해쉬 테이블을 기반으로 구성된다. 쿼드 트리(17)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. *QTH*에서의 쿼드 트리는 히스토그램의 버킷 생성을 위한 분할 알고리즘을 제공하며, 또한 객체의 갱신 정보를 히스토그램에 반영하는 갱신 알고리즘을 제공한다. 해쉬 테이블은 쿼드 트리의 노드에 해당하는 버킷으로 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.

*QTH*는 쿼드 트리 생성, 버킷 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 균등 분할 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 선택을 추정시 영향을 주지 않는다.

*QTH*를 위한 쿼드 트리의 노드 구조는 (numberOfObjects, nodeArea, cPointer, level, flag, objectList)로 구성된다. 여기서 numberOfObjects는 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 버킷을 생성한다. 히스토그램의 버킷은 쿼드 트리의 단말노드 식별자를 해쉬 키로 이용하여 단말 노드에 대응하는 버킷을 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 버킷에 반영할 수 있다.

버킷 생성 후, 각 버킷에는 (nid, MBR, numberOfObjects) 정보를 저장한다. 여기서 nid는 버킷 식별자를 나타내며, MBR은 버킷 내의 공간 경계 영역을 나타내며, numberOfObjects는 버킷 내의 포함된 객체의 수를 의미한다.

쿼드 트리는 노드에서 수용할 수 있는 객체의 수를 초과할 때 공간 분할을 수행함으로써, 객체가 존재하지 않는 빈 노드가 생성된다. 이러한 빈 노드에 버킷을 할당할 경우 불필요한 버킷들이 히스토그램을 구성하게 되므로, 이러한

빈 노드에 대한 버킷을 생성하지 않기 위해 객체의 수가 0인 노드의 경우 버킷을 생성시키지 않는다. 객체의 수가 0인 버킷을 생성하지 않으므로 해서 선택을 추정시 불필요한 연산을 수행하지 않도록 한다.

3.3 현재 질의 선택을 추정

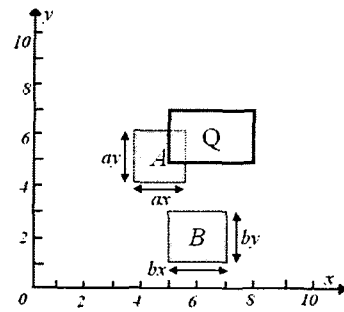
이동객체는 시간에 따라 객체의 위치 및 속도를 변화시킨다. 현재 질의는 이러한 객체의 최근 위치 정보를 기반으로 질의 내에 포함된 객체를 찾는 질의이다. 이러한 현재 질의에 대한 선택을 추정하는 최근 정보를 반영한 히스토그램을 기반으로 추정한다. 현재 질의에 대한 선택을 히스토그램을 구성하는 각 버킷과 질의의 교차 여부를 계산함으로써 추정된다. 이러한 질의와 버킷 간의 기여도 계산 시간을 줄이기 위하여, 고정 그리드 기반의 히스토그램은 prefixsum을 이용하여 상수 시간(constant time)내에 추정할 수 있다. 그러나 균등하지 않은 크기를 갖는 버킷에는 적용할 수 없으므로, 이 논문에서는 [18]의 차원변환을 이용하여 R-Tree의 노드내의 객체 수용량을 추정하는 방식을 이용한다. 이 기법은 QTH를 기반으로 선택을 추정할 때 질의와 히스토그램의 모든 버킷들과 교차 여부를 계산하는데 소요되는 시간을 줄이기 위해, 질의 변환을 통한 버킷 필터링 작업을 수행한다.

【정리 1】 질의 Q는 공간 속성 (qRimin, qRimax)을 가지며, 버킷 B는 공간 속성 (bRimin, bRimax)를 갖는다. 질의 Q가 주어지면, 변환된 질의 Q'를 생성한다. 변환된 질의 Q'의 공간 정보는 (i) q'Rimin = qRimin - (bRimax - bRimin) (ii) q'Rimax = qRimax와 같다. 만일 변환된 질의 Q'이 버킷 B의 좌측 하단 코너 포인트 bRimin를 포함한다면, 질의 Q는 버킷 B와 교차한다.

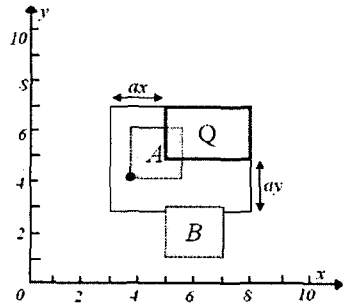
▶ 증명 : i 차원 상의 질의 Q와 버킷 B가 교차한다는 것은 $\max\{qRimin, bRimin\} \leq \min\{qRimax, bRimax\}$ 이며, 질의가 버킷의 최소 포인트 위치 bRimin를 포함하면 이 포인트는 질의와 교차한다. 즉 $qRimin \leq bRimin \leq qRimax$ 이면 질의와 포인트는 교차한다. 이 때 길이 Δx 를 $bRimax - bRimin$ 라 할때, 각 항에 Δx 를 더하면 $qRimin \leq bRimin + \Delta x \leq qRimax + \Delta x$ 가 되며, 이 부등식은 $qRimin - \Delta x \leq bRimin \leq qRimax$ 와 같이 정리된다. Δx 가 $bRimax - bRimin$ 이므로, 부등식은 $qRimin - (bRimax - bRimin) \leq$

$bRimin \leq qRimax$ 와 같이 정리된다. 여기서 $qRimin - (bRimax - bRimin)$ 는 q'Rimin와 같으며, $qRimax$ 는 q'Rimax와 같다. 따라서 $q'Rimin \leq bRimin \leq q'Rimax$ 와 같음을 확인할 수 있다. ■

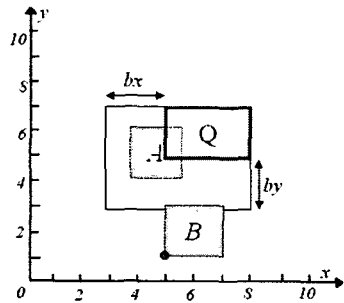
(그림 5)는 변환된 질의를 통한 교차 검사 예를 보여주고 있다. (그림 5(a))에서와 같이 객체 A와 B가 있고, 이들 객체에 대한 질의 Q가 주어지면, 이들 사각형 객체와 질의와의 교차 여부를 포인트와 질의와의 교차 검사로 단순화하여 처리한다.



(a) 질의 및 객체 예



(b) 객체 A에 대한 질의 변환



(c) 객체 B에 대한 질의 변환

그림 5. 질의 변환의 예

Figure 5. The example of Query Transformation

이를 위해 (그림 5(b))와 같이 객체 A에 대한 질의 변환을 수행하기 위해, 객체 A의 각 차원의 길이 a_x 와 a_y 를 질의 Q에 반영하여 질의를 변환시킨다. 변환된 질의는 (그림 5(b))에서와 같이 객체 A의 좌측 코너 포인트를 포함하므로, 원래 질의 Q는 객체 A와 교차한다. 또한 객체 B에 대해서도 동일한 변환과정을 거쳐 변환을 수행할 경우 (그림 5(c))와 같은 결과를 얻을 수 있으며, 그림에서 보여주듯이 변환된 질의가 객체 B의 좌측 하단 코너 포인트를 포함하지 않으므로, 원래 질의 Q는 객체 B와 교차하지 않는다. 위와 같은 차원 변환을 이용한 버킷 필터링을 통하여 질의와 교차 영역을 계산해야 하는 버킷의 수를 줄일 수 있다.

IV. 실험 및 평가

4.1 실험 환경 및 데이터

이 논문에서 제안된 QTH 기법의 성능을 평가하기 위하여 MinSkew 기반 공간 히스토그램, MH 및 R*-Tree 기반의 히스토그램, RH와 다양한 실험 요인들을 변화시키면서 비교 평가 하였다. 모든 기법들은 Intel P4 2.0GHz 시스템의 Window Server 2002 환경에서 C++ 언어로 구현하였다.

히스토그램의 성능은 데이터 분포에 상당히 의존적이다. 이 논문에서는 데이터 셋의 분포를 변화하여 실험 데이터 셋을 생성하기 위해 GSTD(19)를 이용하였다. 실험 데이터 셋은 GSTD를 이용하여 편중된 분포를 갖는 객체 100,000 개를 생성하여 이용하였다. 이때 편중된 데이터 셋의 편중 계수(skewness coefficient)는 0.8로 설정하여 생성하였으며, 생성된 데이터는 공간 범위 $\langle(0,0), (10000, 10000)\rangle$ 내에 존재한다.

히스토그램 기법의 목적은 정확한 선택을 추정하는 것이다. 따라서 정확도는 히스토그램의 성능을 평가하기 위한 중요한 요소가 된다. 이 논문에서 제안된 기법의 정확도를 평가하기 위하여, 저장 공간 및 질의의 크기를 변화시켜가며 평가하였다. 또한 이동객체의 경우 빈번한 위치 변화가 발생하므로, 데이터의 갱신에 따른 정확도를 평가하기 위해, 데이터의 갱신 비율을 전체 데이터 셋의 10~40%까지 변화시켜 평가하였다.

실험 결과인 선택을 추정치의 정확도를 평가하기 위해서 식(2)와 같이 정의된 평균 상대 에러 E_r 를 사용하였다.

$$E_r = \frac{\left(\sum_{q_i=1}^{q_i=n} |S(q_i) - \mathcal{S}(q_i)| \right)}{\sum_{q_i=1}^{q_i=n} \mathcal{S}(q_i)} \times 100 \dots\dots\dots (식 2)$$

식(2)에서 q_i 는 i 번째 질의를 의미하며, 실험에서는 1,000 개의 질의에 대한 평균 에러로 정확도를 평가하였다. $S(q_i)$ 는 질의 q_i 에 대한 실제 선택을 나타내며, $\mathcal{S}'(q_i)$ 는 히스토그램을 이용하여 추정한 선택을 의미한다.

4.2 성능 분석

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 버킷들로 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. 또한 버킷 필터링에 따른 추정 시간을 평가하기 위한 실험도 수행하였다.

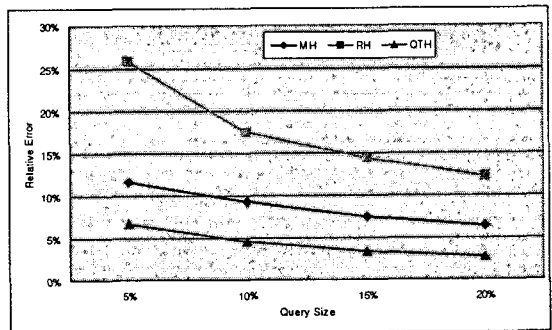


그림 6. 질의 크기에 따른 정확도
Figure 6. Accuracy according to Query Size

(그림 6)은 5%~20% 질의에 대한 평균 추정 결과이다. 실험결과는 모든 기법들이 질의의 크기가 증가할수록 선택을 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 버킷을 할당하여 히스토그램을 생성하게 되므로, 버킷 사이의 중복이 발생하므로, 선택을 추정시 교차하는 버킷에 대한 선택율을 중복 계산하게 되는 문제를 발생시킨다. 따라서 추정 결과는 실제 선택율과 많은 에러를 발생시키게 된다.

(그림 7)은 저장공간 크기에 따른 정확도 실험 결과를 보여주고 있다. 실험 결과는 모든 기법이 버킷의 수가 증가할수록 정확도가 증가하는 경향을 보이고 있다. 이는 버킷의 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 버킷을 생성하지 않으므로 다른 기법에 비해 데이터 분포를 반영할 수 있는 버킷들을 많이 확보할 수 있기 때문이다.

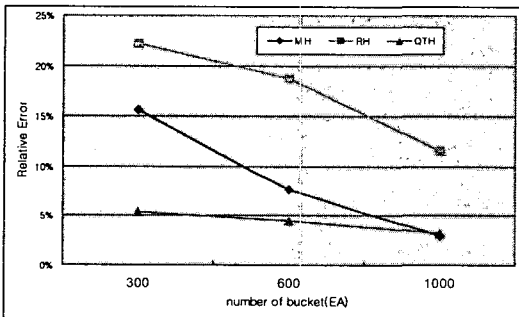


그림 7. 저장공간 크기에 따른 정확도
Figure 7. Accuracy according to Storage Size

(그림 8)은 이동객체의 갱신에 따른 히스토그램 성능의 평가 결과를 보여주고 있다. 이 실험에서는 갱신에 따른 정확도를 평가하기 위해, 전체 데이터의 10~40%까지 갱신된 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. 히스토그램을 빈번히 재생성하면 데이터의 분포를 반영한 요약 정보를 생성할 수 있지만 히스토그램을 생성하기 위해서는 데이터베이스 전체를 다시 스캔해야 하므로 많은 시간적 오버헤드를 초래하게 된다.

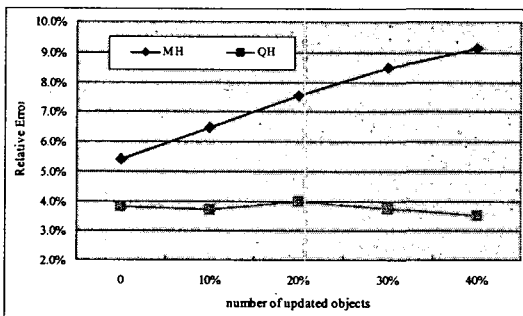


그림 8. 데이터 갱신에 따른 정확도
Figure 8. Accuracy according to data update

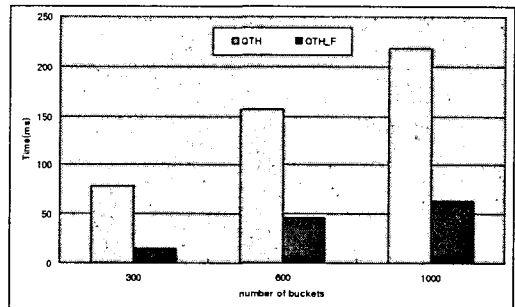


그림 9. 버킷 필터링에 따른 추정 시간
Figure 9. Estimation Time according to Bucket Filtering

(그림 9)는 버킷 필터링 기법을 적용한 후의 추정시간 결과를 나타내고 있다. 실험결과는 버킷의 수가 증가할수록 추정시간이 오래 길어짐을 보여주고 있으며, 또한 필터링을 적용한 기법 QTH_F이 필터링을 적용하지 않은 기법 QTH에 비해 약 3배정도 빠르게 추정할 수 있음을 보여주었다. 이러한 이유는 버킷을 사각형 객체라고 할 경우, 선택을 추정하는 사각형 객체와 사각형 질의 윈도우간의 교차 비율 계산으로 볼 수 있고, 이러한 교차 비율 계산은 두 사각형이 Overlap, Contain 그리고 Containedby 관계 여부를 계산하기 때문에 많은 시간이 걸리게 되기 때문이다.

V. 결론

선택을 추정은 질의 최적화를 위한 기법중의 하나이다. 이동객체의 질의에 대한 기존 선택을 추정 기법은 시간에 따른 이동객체의 위치 변화를 요약 정보에 반영하지 못하며, 또한 기존 공간 요약 정보를 확장하여 이용함으로써 선택을 추정시 많은 에러를 발생시키고 있다. 기존 기법들이 이동객체의 위치 정보 변화를 요약 정보에 반영하기 위해서는 요약 정보를 자주 재생성해야 하며, 그러므로 전체 데이터 베이스를 자주 스캔해야 하는 문제점을 갖고 있다.

따라서 이 논문에서는 이동객체 질의 선택을 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. 또한 해쉬 테이블을 기반으로 유지되므로 많은 양의 색인을 스캔할 필요가 없다. 다양한 실험 결과는 이 논문에서 제안한 쿼드 트리 기반 히스토그램, QTH 기법이 우수한 성능을 가짐을 보였다.

이 논문에서 제안된 기법은 모바일 환경에서 연속적으로 이동하면서 위치 정보를 변경시키는 이동객체의 위치 정보 관리 및 질의 처리를 수행하기 위한 기법이다. 따라서 제안 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] 이기영, 노경택, "공간 데이터 변환 시스템의 설계 및 구현", 한국컴퓨터정보학회논문지, 제8권 제7호, pp.41-46, 2003년 12월.
- [2] 정규장, "웹 기반 공간데이터 공통 컴포넌트 설계 기법", 한국컴퓨터정보학회논문지, 제9권 제1호, pp.31-36, 2004년 3월.
- [3] Y. E. Ioannidis, "Query Optimization", ACM Computing Surveys, Vol.28, No.1, pp.121-123, Mar. 1996.
- [4] Y. E. Ioannidis, and V. Poosala, "Histogram-Based Solutions to Diverse Database Estimation Problems", IEEE Data Engineering Bulletin, Vol.18, No.3, pp.10-18, Sep. 1995.
- [5] V. Poosala, P. J. Haas, and Y. E. Ioannidis, "Improved Histograms for Selectivity Estimation of Range Predicates", ACM SIGMOD, pp.294-305, Jun. 1996.
- [6] 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.
- [7] S. Acharya, V. Poosala, and S. Ramaswamy, "Selectivity Estimation in Spatial Databases", ACM SIGMOD, pp.13-24, Jun. 1999.
- [8] J. H. Chi, S. H. Kim, and K. H. Ryu, "Spatial Selectivity Estimation using Compressed Histogram Information", APWeb, pp.489-494, Mar. 2005.
- [9] J. Jin, N. An, and A. Sivasubramaniam, "Analyzing Range Queries on Spatial Data", ICDE, pp.525-534, Feb. 2000.
- [10] Y. J. Choi, and C. W. Chung, "Selectivity Estimation for Spatio-Temporal Queries to Moving Objects", ACM SIGMOD, pp.440-451, Jun. 2002.
- [11] M. Hadjieleftheriou, G. Kollios, and V. Tsotras, "Performance Evaluation of Spatio-Temporal Selectivity Estimation Techniques", SSDB, pp.202-211, Jul. 2003.
- [12] Y. Tao, J. Sun, and D. Papadias, "Selectivity Estimation for Predictive Spatio-Temporal Queries", ICDE, pp.417-428, Mar. 2003.
- [13] Q. Zhang, and X. Lin, "Clustering Moving Objects for Spatio-Temporal Selectivity Estimation", ADC, pp.123-130, Jan. 2004.

- [14] H. G. Elmongui, M. F. Mokbel, and W. G. Aref, "Spatio-temporal Histogram", SSID, pp.19-36, Aug. 2005.
- [15] M. Wang, J. S. Vitter, L. Lim, and S. Padmanabhan, "Wavelet-based Cost Estimation for Spatial Queries", SSTD, pp.175-196, Jul. 2001.
- [16] 선종복, 김진덕, 김장수, 홍봉희, "공간질의 최적화를 위한 선택률 추정 방법," 정보과학회 논문지(D), 제 25권 제7호, pp980-995, 1998년 7월.
- [17] H.Samet, "The Design and Analysis of Spatial Data Structures," Addison Wesley Publishing Company, INC., 1994.
- [18] I. Kamel, and C. Faloutsos, "On Packing R-Trees", CIKM, pp.490-499, Nov. 1993.
- [19] 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.

저 자 소 개



지 정 희
 1999년 충주대학교 전자계산학과 졸업
 2001년 충주대학교 대학원
 전자계산학 석사
 2006년 충북대학교 대학원
 전자계산학 박사
 <관심분야> 시공간 DB, Temporal GIS, 스트림 데이터베이스



류 근 호
 1976년 숭실대학교 전자계산학과 졸업
 1980년 연세대학교 공학대학원
 전자계산학 석사
 1988년 연세대학교 대학원
 전자계산학 박사
 1976년 ~ 1986년 육군군수지원사
 전산실(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, 유비쿼터스
 컴퓨팅 및 스트림 데이터,
 지식기반 정보검색시스템,
 데이터 마이닝, 데이터베이스
 보안 및 Bio-Informatics

정 두 영
 1973.2 서강대 졸업
 1987.2 충북대학교 정보통신과 교수
 <관심분야> 데이터베이스, 컴퓨터 망