시공간 데이타베이스의 엔트로피 기반 동적 히스토그램

(Entropy-based Dynamic Histogram for Spatio-temporal Databases)

박 현 규 † 손 진 현 †† 김 명 호 †††
(Hyun Kyoo Park) (Jin Hyun Son) (Myoung Ho Kim)

요 약 질의 최적화에 사용하기 위한 선택도 추정 방법은 히스토그램, 샘플링 그리고 패러미터에 의한 요약 방법 등이 제시되고 있다. 히스토그램을 이용한 선택도 추정은 상용 데이타베이스 시스템에서 가장 보편적으로 사용되는 방법이지만, 이동 객체를 위한 시공간 데이타베이스에서는 데이타 분포가 지속적으로 변화함으로써 기존의 히스토그램 방법을 이용하는 것은 제한이 많게 된다. 특히 미래 질의를 위해서는 데이타 갱신을 반영하는 동적 관리가 가능하며, 정확도를 유지할 수 있는 다른 접근 방법이 필요하다.

따라서 시공간 객체를 위한 선택도 추정 방법은 질의 술어가 요구하는 데이타 분포에 대한 히스토그램이 필요하며, 본 논문에서는 미래의 시공간 영역 질의 술어에 대하여 신속히 히스토그램을 생성할 수 있도록 쌍대성과 한계 분포 방법을 이용한 히스토그램을 제안한다. 쌍대 공간에서 이동 객체에 대한 데이타 시놉시스를 이용하여 구성된 시공간 히스토그램은 이동 궤적의 선형성이 유지하는 시간 동안 정확성을 보장하면서 빠른 시간에 생성이 가능하다. 또한 동적 갱신을 점증적으로 지원함으로써 효율적으로 갱신된 정보를 반영할 수 있고 추정 결과의 정확성을 향상시킬 수 있다.

키워드 : 시공간 데이타베이스, 이동 객체, 선택도 추정, 히스토그램

Abstract Various techniques including histograms, sampling and parametric techniques have been proposed to estimate query result sizes for the query optimization. Histogram-based techniques are the most widely used form for the selectivity estimation in relational database systems. However, in the spatio-temporal databases for the moving objects, the continual changes of the data distribution suffer the direct utilization of the state of the art histogram techniques. Specifically for the future queries, we need another methodology that considers the updated information and keeps the accuracy of the result.

In this paper we propose a novel approach based upon the duality and the marginal distribution to construct a histogram with very little time since the spatio-temporal histogram requires the data distribution defined by query predicates. We use data synopsis method in the dual space to construct spatio-temporal histograms. Our method is robust to changing data distributions during a certain period of time while the objects keep the linear movements. An additional feature of our approach supports the dynamic update incrementally and maintains the accuracy of the estimated result.

Key words : Spatio-temporal Database, Moving Object, Selectivity Estimation, Histogram

· 본 연구는 CEBT 및 2002년 한양대학교 교내연구비의 부분적인 지원을 받았음.
† 종신회원 : 한국과학기술원 전자전산학과 전산학
hkpark@dbserver.kaist.ac.kr
†† 종신회원 : 한양대학교 컴퓨터공학과
jhson@dbserver.kaist.ac.kr
††† 종신회원 : 한국과학기술원 전자전산학과 전산학 교수
mhkim@dbserver.kaist.ac.kr
논문접수 : 2002년 7월 27일
심사완료 : 2002년 10월 31일

1. 서 론

선택도 추정(Selectivity Estimation)은 질의 최적화(Query Optimizer) 등에서 다양하게 사용되며, 최근 위치 기반 서비스(Location-Based Service)가 보편화되면서 시공간 데이타베이스에서도 질의 술어(Predicate)와 관련된 정확하고 효율적인 선택도 추정 방법이 요구되고 있다.

위치 기반 서비스를 위한 시공간 데이타베이스는 시간에 따라 지속적으로 변화하는 이동 객체의 위치 정보를 다루고 있으며, 서비스에서 요구되는 미래의 위치 기반 시공간 영역 질의는 이동 객체의 예상 궤적을 이용하여 미래의 위치 정보를 추정하여 검색하는 질의로서 기존의 질의와 구별된다.

히스토그램은 선택도 추정을 위하여 보편적으로 사용되는 방법이지만 데이타 분포의 상당한 변화가 있는 경우 이를 바로 반영하지 못하는 정적인 구조(Static Structure)이다. 따라서 시간에 따라 지속적으로 데이타 분포가 변화하는 위치 정보를 다루는 시공간 데이타베이스에서 선택도 추정을 위하여 히스토그램을 사용하는 것은 결과의 부정확성이 크거나, 추정 과정이 매우 복잡해 지게 된다. 또한 히스토그램을 구성한 이후 위치 정보의 갱신을 적시에 반영하기 어렵다.

본 논문에서는 질의 술어가 요구하는 이동 객체에 대한 위치 정보 분포를 다차원 히스토그램으로 신속하게 생성하여 효율적인 선택도 추정이 가능한 동적 관리 기법을 제시하였다. 이동 객체의 위치 정보는 특정 시간의 위치와 이동 속도 정보를 저장함으로써 궤적 정보를 보다 효율적으로 데이타베이스에 저장할 수 있으며 [1, 2], 논문에서는 함수로 저장된 위치 정보를 쌍대 공간(Dual Space)에서 만들어진 시놉시스(Synopsis)로 구성함으로써 질의 술어에 따른 히스토그램을 실시간으로 생성하는 방법을 제시한다. 이 방법은 또한 데이타베이스에 갱신이 발생한 경우 이를 시놉시스에 증감(Incremental) 형태로 반영함으로써 히스토그램 갱신의 오버헤드를 최소화시키고, 보다 정확한 질의 결과를 얻을 수 있는 장점이 있다.

논문의 구성은 다음과 같다. 2장에서는 문제 정의와 시공간 데이타의 특성에 따른 연구 동기를 기술하고, 3장은 본 연구와 관련된 기존 연구, 4장은 시놉시스를 이용한 히스토그램의 생성과 선택도 추정 방법 그리고 데이타 갱신을 반영하는 동적 관리 방법을 기술하였다. 그리고 5장은 제시된 방법의 효율성을 분석하고, 6장에서는 결론 및 향후 연구 과제를 제시하였다.

2. 문제 정의 및 연구 동기

미래 위치 기반 질의에 대한 선택도 추정은 데이타베이스의 저장된 데이타를 이용하여 예상되는 궤적 정보를 대상으로 하는 시공간 질의이다. 즉, $O = \{o_1, o_2, ..., o_n\}$가 n개의 궤적으로 이루어진 객체들의 집합일 때 시간 변화에 따른 객체의 위치는 궤적 상의 좌표가 되고, 시간 t에서의 객체의 위치는 $o_i(t)$로 표시한다. 이동 객체의 궤적은 불확실성을 포함하여 구간 단위(Piece-wise) 선형 함수로 가정하며[2, 3, 4], 위치 정보는 객체의 위치와 이동 속도에 의하여 x좌표는 시간에 따른 함수 $x_i(t) = a_i + b_it$로 표현 되고, y좌표도 동일한 방법으로 표현된다.

그러므로 논문에서 다루고자 하는 질의 술어(Predicate)에 대한 선택도 추정을 위해서는 이동 객체의 현재 위치와 이동 속도를 이용하여 요구되는 질의 시간에서 근사 질의 결과를 신속하게 얻을 수 있는 히스토그램이 요구된다. 그러나 미래 시간의 이동 객체에 대한 대한 선택도 추정은 질의 술어에 대한 다차원 히스토그램이 필요하지만 질의가 이루어지는 시간에 데이타베이스를 검색하여 히스토그램을 생성하는 것은 비현실적이며, 따라서 기존의 히스토그램 방법을 직접 이용하는 것은 곤란하다.

즉, 히스토그램을 이용한 선택도 추정은 그림 1-(a)와 같이 객체의 위치를 시간 t_0에서 구성된 1차원 히스토그램을 이용하여 미래 시간의 위치 정보를 추정하는 방법을 사용하거나[5], 그림 1-(b),(c)와 같이 요구되는 현재 또는 미래 시간의 객체 위치 분포를 다차원 히스토그램을 이용하여 추정하는 방법을 사용할 수 있다[6].

그러나 기존의 1차원 히스토그램을 이용하는 방법은 버킷 B_i의 정보가 t_0에서 구성됨으로써 x축에서 각 버킷에 존재하는 객체의 수가 시간에 따라 변화하여 결과의 정확도(Accuracy)가 떨어지고, 질의 처리가 복잡해지는 단점이 있다. 즉, $B_i(t_0)$에 해당하는 객체는 시간에 따른 x축 상의 변화를 포함해야 하므로 t_0와 질의 시간 차이가 클수록 버킷이 표현하는 위치 분포 정보는 정확도가 떨어지며, 이를 극복하기 위해서는 히스토그램의 갱신 주기가 대단히 짧아야 한다. 반면에 다차원 히스토그램을 이용하는 방법은 공간 데이타베이스에서 다루어진 최적의 히스토그램을 이용할 수 있지만 모든 시간에 대한 히스토그램을 미리 생성하는 것은 불가능하다.

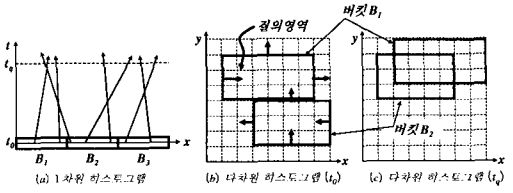


그림 1 히스토그램에 의한 선택도 추정

위치 정보를 기반으로 하는 시공간 데이타베이스의 응용 영역은 도심 지역의 교통 관리(Traffic Management)와 같이 불규칙적으로 움직이는 많은 차량을 대상으로 하는 경우뿐만 아니라 해상에서 움직이는 함대 또는 기

계화 부대와 같은 이동 객체를 다루는 군사 목적의 지휘통제 시스템 등이 있을 수 있다. 그러므로 선택도 추정은 질의 최적화뿐만 아니라 질의에 대한 근사 결과(Approximate Answers)를 빠른 시간에 요구하는 군사 목적의 시스템과 같은 응용 서비스에서는 갱신된 정보가 질의 결과에 반드시 반영되어야 하는 경우가 있을 수 있다.

그러므로 본 논문에서는 기존의 히스토그램을 이용한 효율적인 분할 또는 질의 처리 방법보다는 정확도가 높고 질의 처리가 상대적으로 용이한 질의 술어에 해당하는 다차원 히스토그램을 신속하게 생성하는 방법에 중점을 두고 있다. 히스토그램은 버킷 내에서 객체는 균일 확산(Spread)과 균일 빈도(Frequency)로 구성된 것으로 가정하며[7], 본 논문에서는 이러한 가정 하에서 선형 함수로 표현되는 이동 궤적에 대한 히스토그램을 생성할 수 있도록 한다. 또한 제시된 방법은 궤적의 변화가 발생하는 경우에도 히스토그램의 버킷 정보에 갱신된 데이타를 실시간에 반영토록 함으로써 동적으로 데이타 분포를 보다 정확히 표현할 수 있는 관리 방법을 제시한다.

3. 관련 연구

질의 최적화를 위한 선택도 추정 및 질의 처리 근사값을 효율적으로 처리하는 방법으로 히스토그램[7, 8, 9], 웨이블릿(Wavelet)[10, 11] 등이 제시되고 있다. 히스토그램은 정렬(Sort) 및 원천 패러미터(Source Parameter), 분할 범주(Partition Class), 분할 조건(Partition Constraint)에 따라 구분할 수 있으며[9], 기존 연구 결과는 다양한 히스토그램 구성 방법 가운데 *Min-Skew* 히스토그램[7]이 상대적으로 정확도가 높고 저장 공간을 작게 차지함을 제시하였다. 그러나 전체 데이타를 읽어 들여 구성하는 *Min-Skew* 히스토그램을 지속적으로 변화하는 데이타에 적용하기는 곤란하다.

본 논문에서는 질의 술어가 요구하는 히스토그램을 생성할 수 있도록 쌍대 공간에서 데이타 시놉시스를 이용하는 방법을 사용하며, 미래 질의를 처리하기 위하여 쌍대 공간을 이용하여 궤적을 저장, 색인하는 방법은 [1, 2]등에서 제시되었다. 쌍대 공간에서 점으로 표현된 궤적 정보를 요약하는 시놉시스는 패러미터(Parameter) 기반과 비패러미터(Non-Parameter) 기반으로 구분되며, 패러미터에 의한 방법은 통상 정규 분포의 z값 등을 이용한 방법 등이 있고, 비패러미터 방법은 트리(Tree)를 이용하는 방법 등이 있다. 시놉시스로 데이타를 표현하기 위해서는 데이타 영역을 중첩하거나 중첩하지 않는 영역으로 분할해야 하며, 다차원 공간을 분할하는 방법은 특수한 경우를 제외하고 대부분의 다차원 분할에 관한 최적화 문제는 NP-Hard로 알려져 있다[13].

시공간 데이타베이스에서 다루는 객체는 시간에 따라 지속적으로 변화하는 데이타이므로 갱신 내용을 동적으로 반영할 수 있어야 한다. 동적 히스토그램 관리 방법으로 동적 히스토그램[6], 웨이블릿[11]등과 같은 방법을 고려할 수 있으나, 이들은 전체 객체 중에서 일부의 갱신이 있는 경우에 적합하며 갱신 결과가 전체 구조에 영향을 미치는 경우에는 동적 관리의 복잡도가 매우 증가하게 된다[6, 12].

보다 작은 저장 공간을 이용하여 전체 데이타 분포를 표현하는 연구는 다양한 분야에서 요구되고 있으며, 웨이블릿을 이용하여 데이타 시놉시스를 생성하는 방법에 대한 연구는 [11]이 있다. 그러나 본 논문에서 제시하는 시놉시스는 이전 연구에서 거의 다루지 않고 있는 동적 관리 부분을 포함하여 효과적으로 히스토그램을 생성할 수 있는 방법으로 기존 연구와 구별된다.

4. 엔트로피 기반 시공간 히스토그램

위치 정보 기반 미래 질의에 대한 선택도 추정을 위해서는 질의 술어가 요구하는 시간의 다차원 히스토그램을 효율적으로 생성할 수 있어야 한다. 그러나 선택도 추정을 위하여 질의가 이루어질 때마다 전체 데이타 분포를 읽을 수 없으므로, 본 논문에서는 이동 객체의 궤적이 갖는 선형성을 기반으로 쌍대 공간에서 시놉시스를 구성함으로써 히스토그램을 동적으로 생성할 수 있는 방법을 제시한다.

정의 1. 엔트로피

β개의 버킷으로 이루어진 히스토그램의 버킷에 존재하는 객체의 수를 확률 분포 p로 표현한 것을 엔트로피라 하며, 엔트로피는 $(p_i \log p_i)$ 로 표시되고 전체 버킷이 표현하는 엔트로피는 $\sum_{i=1}^{\beta} p_i \log p_i$가 된다. □

확률 및 정보 이론에서 다양하게 사용되는 엔트로피는 확률 분포에 의한 정보의 양으로 히스토그램에서 각 버킷에 존재하는 객체의 수를 표현하는데 적용할 수 있다. β개의 버킷으로 이루어진 시공간 히스토그램의 버킷 B_i 에 존재하는 객체의 수는 n_i이므로 전체 객체의 수는 $\sum_{i=1}^{\beta} n_i$가 되며, 시공간 히스토그램에 의한 선택도 추정은 질의 영역과 겹치는 버킷 내에 존재하는 객체의 수가 된다. 따라서 히스토그램의 정보는 버킷들로 구성된 각 행과 열의 집계 합을 한계 분포(Marginal Distribution)로 사상시킬 수 있으며, 한계 분포의 값을 구성하는 각 버킷의 정보를 정의 1과 같이 엔트로피를 이용하여 표현 가능하다.

4.1 쌍대 공간에서의 시놉시스

이동 객체의 위치 정보는 시간 변화에 따라 xy공간 상의 점객체로 표현되며, 객체의 이동이 선형(Linear) 일 때 예상 궤적은 시간에 따른 선형 함수 $\{F_x(t), F_y(t)\}$의 형태로 표현된다. 선형 함수의 쌍대성(Duality)을 이용하면 주(Primal) 공간에서 x축에 대한 함수 $x=vt+x'$로 표현되는 궤적을 쌍대 공간에서 점$(v+x')$으로 표현할 수 있으며, y축에 대해서도 동일한 과정을 통하여 3차원 공간에서 궤적은 2개의 2차원 점 객체로 표현할 수 있다[1, 2, 14]. 쌍대 공간에서는 객체의 이동이 선형일 때 점으로 표현된 정보의 변화가 없으므로 이동 궤적의 저장과 질의 처리에 보다 용이한 방법을 제공할 수 있다.

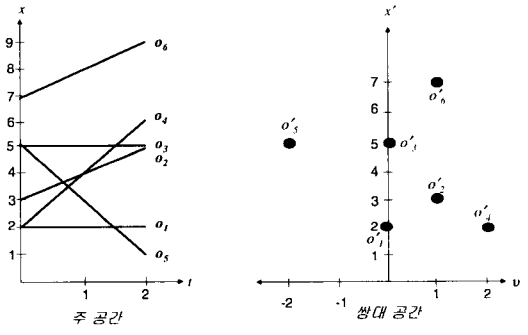


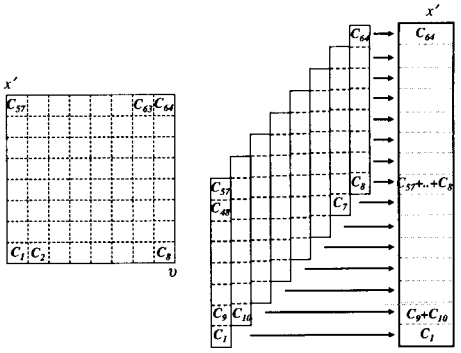
그림 2 주 공간과 쌍대 공간에서 궤적 표현

예로서 그림 2는 주 공간과 쌍대 공간에서 객체의 예상 궤적을 표현한 것으로써, 주 공간 xt에서 궤적은 쌍대 공간 x((에서 점으로 표현된다. 그러므로 그림 2의 주 공간에서 객체 o_i(i=1..6)의 궤적은 쌍대 공간에서 o'_i (i=1..6)의 점으로 표현된다.

정의 2. 시놉시스

시놉시스 $S(t)$는 데이타 분포를 그리드(Grid)로 분할하여 셀(Cell)로 이루어진 2차원 배열이며, 그리드의 범위와 범위 내에 존재하는 객체 수, 평균 속도를 정보로 저장한다. □

본 논문에서는 쌍대 공간에서 시간 t0 일 때 x축과 y축에 대하여 데이타 분포를 시놉시스로 각각 생성하며, 그림 3-(a)는 x축에 대한 시놉시스를 구성한 예제이다. 그림 3-(a)와 같이 t0에서 구성된 시놉시스는 정방형의 공간으로 분할되며, 질의 술어가 요구하는 히스토그램을 생성하기 위해서는 시놉시스를 그림 3-(b)와 같이 조정한다. 그림 4는 시놉시스로부터 히스토그램을 구성하기 위한 한계 분포를 구하는 알고리즘을 나타내었다.



(a) 그리드 분할 시놉시스(t_0) (b) 시놉시스 조정과 집계 합 배열 (t_q)

그림 3 쌍대 공간에서 공간 분할에 의한 시놉스와 질의 술어에 따른 한계 분포

한계 분포는 시놉시스의 집계 합(Aggregated Sum) 배열로 구할 수 있으며, 질의 술어에 의한 tq에서 시놉시스 가장 좌측 셀들의 행(Column)은 (x0′-((tq-t0))만큼 아래로 이동하고, 가장 우측 셀들의 행은 (x0′+((tq - t0))만큼 위로 이동한다. 즉, (축의 중앙을 중심으로 좌측 절반에 해당하는 행들은 질의 시간과 시놉시스 구성 시간의 차이만큼 아래로 조정되며, 우측 절반의 행들은 위로 조정되어 그림 3-(b)와 같이 한계 분포를 구성한다. 쌍대 공간에서 속도에 의하여 조정된 행들을 이용하여 시간 tq에서 구성된 집계 합 배열은 질의 시간에 해당하는 쌍대 공간 x′축에서의 분할된 그리드 범위와 객체의 수를 정보로 하는 1차원 배열이 된다.

Algorithm Marginal_Distribution_by_Query_Predicates
Input: Query Predicates and Synopsis
Output: Marginal Distribution of each axis
1. Shift down left half columns of synopsis based on the query predicate
2. Shift up right half columns of synopsis based on the query predicate
3. Aggregate cell information to 1-dimensional array as an output
4. Count and store the number of entries at each row to array
5. return Marginal_Distribution

그림 4 히스토그램을 위한 한계 분포 추출 알고리즘

예제 1. 그림 3에서 시놉시스는 $t=0$에서 생성되고, 질의 술어가 t_q=10인 히스토그램을 생성할 것을 요구한다고 하면, 한계 분포 x'는 다음과 같다. 셀 C1이 구간 (10000,

11000), 평균 속도 10일 때 한계 분포의 최하단 배열의 값은 평균 속도와 시간 구간의 차를 곱한 100을 하향 조정한 (9900, 10900) 구간에서 객체의 수를 표현하게 된다. □

4.2 질의 술어 기반 히스토그램 생성

시뇹시스를 이용하여 질의 술어에 해당하는 시간에 대한 객체의 분포를 표현한 다차원 히스토그램은 실제 데이타 분포를 이용하여 생성되는 히스토그램을 최대한 근사화 시킬 수 있어야 한다. 이를 위하여 본 논문에서 사용되는 히스토그램은 다음과 같이 정의한다.

정의 3. 시공간 히스토그램

질의 술어에 대한 히스토그램은 중첩되지 않는 β개의 버킷으로 구성된 2차원 히스토그램이며, 각 버킷 B_i의 정보는 공간 범위 $[x_l, x_u]$, $[y_l, y_u]$과 객체의 수 n으로 이루어진 집합 $\{x_l, x_u, y_l, y_u, n\}$으로 구성되고, 객체는 버킷 내에서 균일 분포한다. □

논문에서 제시하고자 하는 시공간 히스토그램은 시뇹시스 S로부터 얻어진 한계 분포를 이용하며, $P(A)$를 2차원 주 공간에 대한 확률 분포라고 하면, 히스토그램을 구성하는 문제는 다음과 같이 정형화시킬 수 있다.

문제

다음을 만족하는 2차원 확률 분포 P에 대한 엔트로피 $E(P)$의 최대화

$P(A) \geq 0$, $\sum P(A) = 1$, 그리고

$\forall_i \in S$, $\sum_{j \in S-i} P(j) = P(i)$, $P(j)$는 $j \in S$인 P의 한계 분포 □

엔트로피를 최대화하는 방법은 시뇹시스를 이용하여 얻어진 x', y'에 대한 1차원 배열을 이용하여 히스토그램의 확률 분포 $P(A)$를 최대화함으로써 해결할 수 있다. 엔트로피의 척도는 두개의 확률 분포 사이의 유사도(Likelyhood)이므로, 문제 정의와 같은 엔트로피 최대화는 실제 분포와 한계치(Marginals)로부터 얻어진 확률 분포 사이의 차이를 최소화시키는 문제와 동일하다. 그러므로 히스토그램의 버킷 $b = \{B_1, B_2, ..., B_\beta\}$의 정보를 표현하는 객체의 분포 확률이 $p_i \geq 0$그리고 $\sum_{i=1}^{\beta} p_i = 1$를 만족하는 실제 확률 분포 $p(a,M)$과 추정된 확률 분포 $p(a, \hat{M})$사이의 거리(Distance)를 최소화하는 $\hat{M} = (P_1, \cdots, P_\beta)$을 구하는 문제와 동일하다.

정의 4. Kullback-Leibler 거리

Kullback-Leibler 거리는 두가지 확률 분포 함수 사이의 차이이며 다음과 같다. □

$$D(p(A;\hat{M}) \| p(A;M)) = \int p(a;\hat{M}) \log \frac{p(a;\hat{M})}{p(a;M)} da$$

k개의 버킷으로 이루어진 히스토그램으로 표현된 객체

Algorithm Spatiotemporal_Histogram_Construction
Input: Marginal Distribution for x and y axis
Output: Histogram for query predicate
1. Let values of output array to 1; i =0; j = 1; t = 1;
2. do {
3. $\hat{V}^t_{a1,a2} = \hat{V}^{t-1}_{a1,a2} \frac{p_i(a_j)}{\hat{p}_{i+1}(a_j)}$;
4. i++; t++;
5. if j == 2 then j = 1 else j = 2;
6. } while sum of the output array is the same of the marginal distribution

그림 5 동적 히스토그램 생성 알고리즘

수에 대한 정보를 V라고 하면, 실제 확률 분포와 추정 확률 분포의 차이를 *Kullback-Leibler* 거리로 표현할 수 있으며, 시뇹시스로부터 히스토그램을 생성하는 과정은 알고리즘 그림 5와 같다. *Kullback-Leibler* 거리는 객체의 수가 충분히 큰 데이타 집합일 때 추정 패러미터는 실제 패러미터에 수렴하므로[15], 시뇹시스로부터 얻어진 한계 분포를 이용하여 이산 확률 분포에 의한 *Kullback-Leibler* 거리는 (1)과 같다[16].

$$D(\hat{p} \| p) = \sum_{i=1}^{\beta} \hat{p}_i \log \frac{\hat{p}_i}{p_i} =$$

$$\sum_{i=1}^{\beta} \hat{p}_i \log \hat{p}_i - \sum_{i=1}^{\beta} \hat{p}_i \log p_i \qquad (1)$$

따라서 공식 (1)의 정보는 히스토그램의 각 버킷이 표현하는 정보로 사상시켜 한계 분포를 구성하는 엔트로피를 구하는 과정이 된다.

그림 6은 알고리즘에 대한 예제로서 4×4로 이루어진 히스토그램 생성 과정을 표시하였다. 시뇹시스로부터 얻어진 x', y' 가 그림 6-(a)와 같을 때, 히스토그램은 그림 6-(b)와 같이 모든 버킷의 값을 1로 초기화한다. 각 버킷은 x축과 y축을 분할한 공간이므로 속성 a_1과 a_2에 대한 엔트로피 $p(a_1)$과 $p(a_2)$를 최대화 시키는 순환 알고리즘을 수행한다.

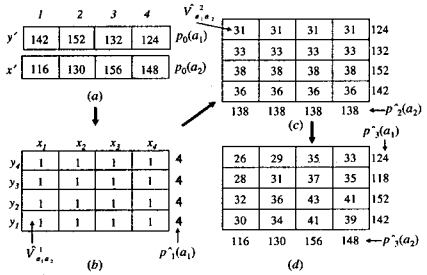


그림 6 질의 술어에 따른 시뇹시스로부터 다차원 히스토그램의 생성

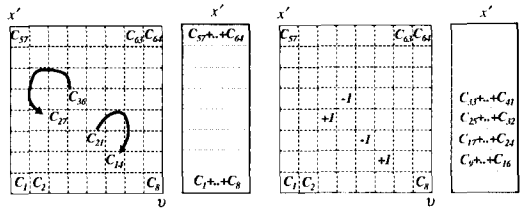
그림 6-(b)의 값 $\hat{p}_1(a_1)$은 각 열의 버킷들의 합이며, 다음 단계인 그림 6-(c)와 같이 $V^3_{a1,a2} = V^2_{a1,a2}\frac{p_0(a_1)}{\hat{p}_1(a_1)}$를 각 버킷의 값으로 설정하여 행의 값 $\hat{p}_2(a_2)$를 얻는다. 최종적인 버킷의 값은 그림 6-(d)와 같이 $V^3_{a1,a2} = V^2_{a1,a2}\frac{p_0(a_1)}{\hat{p}_2(a_1)}$가 된다. 이 때 알고리즘은 추정 값 $\hat{p}_3(a_1), \hat{p}_3(a_2)$가 한계 분포와 동일할 때 종료된다.

4.3 동적 갱신의 히스토그램 반영

동적 히스토그램에 관련된 연구는 갱신된 레코드에 의한 데이타 분포 변화를 히스토그램에 반영할 수 있도록 최소한의 오버헤드를 포함하여 버킷 정보들을 재배치(Rebalancing)하는 부분에 중점을 두고 있다[6]. 그러나 이러한 방법은 상대적으로 적은 수의 갱신이 이루어지거나 재배치가 전체 구조에 큰 영향을 미치지 않는 경우에 사용할 수 있다.

시뮬시스로 표현된 모든 객체들의 궤적은 속도와 방향을 일정 시간 동안 유지하는 것을 가정하지만 실제 환경에서는 도로 상황 또는 목적지 변경 등으로 궤적의 변화가 발생할 수 있고, 시공간 데이타베이스의 미래 질의는 수시로 변화하는 객체의 궤적을 반영할 수 있어야 선택도 추정의 정확도를 높일 수 있다.

히스토그램의 변화에 영향을 미치는 사건은 객체의 삽입, 삭제뿐만 아니라 기존 객체 궤적의 변화로 구분할 수 있다. 따라서 본 논문에서 제안된 방법은 모든 경우의 갱신에 대하여 시뮬시스의 해당 되는 셀의 정보에 삽입은 (+1), 삭제는 (-1)을 반영함으로써 동적 히스토그램 갱신이 가능하다.



(a) 데이터 갱신 반영 (b) 갱신된 시뮬시스 생성

그림 7 데이타 갱신에 의한 점증적 시뮬시스 갱신

그림 7은 시뮬시스에 갱신 내용을 반영한 예제이며, 그림 7-(a)와 같이 C36 셀에 해당하는 객체 궤적이 C27로 갱신된 경우와 C21 셀에 존재하는 객체가 C14로 갱신된 경우이다. 이러한 갱신 내용을 시뮬시스에 반영하는 과정은 그림 7-(b)와 같이 과거 데이타는 삭제되고 갱신된 데이타는 삽입하는 과정과 동일하므로 해당 셀의 정보에서 객체 수를 증가 또는 감소시킴으로써 수행된다.

따라서 히스토그램을 위한 집계 합 배열을 구성하는 과정이나 히스토그램 생성에 영향를 미치지 않고 선형 시간에 갱신 내용을 시뮬시스에 반영할 수 있으므로 시스템의 성능에 영향을 미치지 않는 범위 내에서 주기적으로 일정 범위 이상의 갱신이 발생한 경우 효율적으로 시뮬시스에 갱신 내용을 반영할 수 있다.

5. 히스토그램의 성능 및 효율성 분석

이동 객체 위치 정보에 대한 히스토그램은 쌍대 공간에서 시뮬시스의 생성과 시뮬시스를 이용하여 히스토그램을 구성하는 과정에 대한 복잡도로 구분하여 평가할 수 있다. 즉, N개의 객체를 이진 검색을 이용하여 n개의 그리드로 분할된 쌍대공간에서 해당되는 셀 정보를 갱신하는 복잡도는 개별 객체에 대해서는 $O(\log n)$이고, 전체 구성 복잡도는 $O(N \log n)$이 된다.

이 경우 시뮬시스를 위하여 정확도와 효율성을 유지하는 범위에서 소요되는 셀의 숫자는 수백 이하이므로 [9], 시뮬시스를 이용하여 집계 합 배열을 구성하는 시간은 $O(n)$, 동적 갱신 관리를 위한 복잡도는 거의 상수 시간(Constant Time)에 처리될 수 있다.

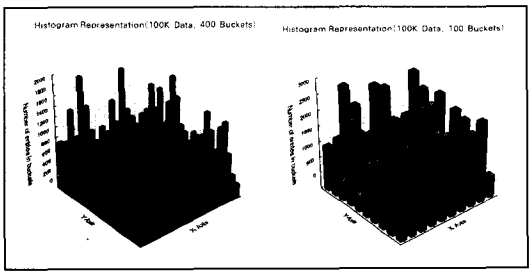


그림 8 시뮬시스로부터 구성된 히스토그램

질의 술어에 의한 다차원 히스토그램을 생성하는 과정은 xy 2차원 공간에서 각 버킷이 가지는 정보를 추출하는 것이며, 2차원 배열의 β버킷이 가질 수 있는 집합은 $\binom{2}{\beta}$이므로 히스토그램의 멱집합(Power Set)으로부터 한계 분포를 이용한 최종 히스토그램을 생성하는 복잡도는 그림 5의 알고리즘으로부터 최대 $\left[\binom{2}{\beta}\right]$의 순환이 일어나게 된다. 본 논문에서 제시된 방법이 실제 환경에서 이동 객체를 위한 시공간 히스토그램으로 적용이 가능한

지 검증하기 위하여 실험을 통한 효율성과 정확성을 분석하고자 한다.

실험에서 사용된 데이타는 일반적인 도시 환경에서 다양한 이동 형태를 묘사할 수 있는 *IBM citySimulator*[1] 를 이용하여 생성하였다. 그림 8은 논문에서 제시한 방법에 의하여 생성된 시공간 히스토그램의 예로서 *xy*공간에서 이동 객체 100,000개의 위치 정보에 대하여 버킷의 수를 100과 400으로 각각 설정하여 질의 술어에서 요구되는 시간에 구성된 히스토그램이다.

본 논문에서 제시하는 히스토그램의 구성은 복잡도 분석과 같이 실제 구성 시간이 수 밀리초 이내에 수행되므로 매우 신속한 결과를 얻을 수 있으나, 히스토그램이 의미를 가지기 위해서는 효율성뿐만 아니라 결과의 정확도를 제공할 수 있어야 한다. 히스토그램의 정확도는 절대 오차(Absolute Error)와 상대 오차(Relative Error)로 평가할 수 있으며[7, 8], 시공간 질의 q_i에 의한 결과 θi와 히스토그램에 의한 추정 결과 θ'_i로부터 절대 오차 e^{abs}와 상대 오차 e^{rel}는 다음과 같이 계산된다.

$$e_i^{abs} = |\theta_1 - \theta'_i|, \quad e_i^{re} = e\frac{abs}{i}/\theta_i = \frac{|\theta_1 - \theta'_i|}{\theta_i} \qquad (2)$$

그리고 질의 집합 Q 에 대하여, 평균 절대 오차 Eabs 와 평균 상대 오차 Erel 는 다음과 같다.

$$E^{abs} = \frac{1}{Q}\sum_{i=1}^{Q} e_i^{abs}, \quad E^{rel} = \frac{1}{Q}\sum_{i=1}^{Q} e_i^{rel} \qquad (3)$$

정확도는 버킷의 수와 데이타의 크기, 그리고 데이타 분포 특성에 의하여 영향을 받는다. 실험에서는 도시환경의 특성을 반영할 수 있도록 48개의 도로, 71개의 빌딩, 1개의 공원으로 이루어진 공간에서 10만~50만 이동 객체를 생성하고 수행된 질의에 대한 정확한 결과를 전체 데이타베이스 검색을 통하여 얻은 후 실험을 수행하였다.

히스토그램에 의한 정확도는 데이타 크기와 버킷의 수의 변화에 따라 그림 9와 같이 나타나며, 20만 이상의 객체에 대해서는 버킷의 수가 많은 경우 정확도의 충분한 개선을 기대할 수 있음을 보인다.

또한 질의 영역의 크기와 데이타 크기, 버킷의 수가 선택도 추정의 정확도에 미치는 영향을 그림 10에서 제시하였다. 논문에서 제시한 방법은 데이타 크기가 상대적으로 큰 경우 보다 정확도를 향상시킬 수 있음을 보이며, 시놉시스를 동적으로 관리함으로써 제시된 정확도는 거의 모든 데이타 집합에 대하여 유지할 수 있었다.

1) LBS를 위한 이동 객체 시뮬레이터 IBM Software(http://alphwaworks.ibm.com)

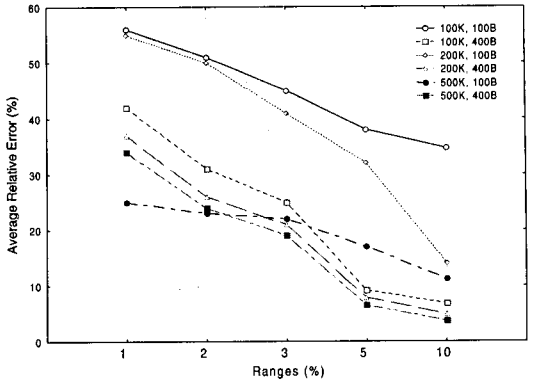


그림 9 버킷과 데이터 크기에 이한 히스토그램 정확도

그리고 실험을 통하여 논문에서 제시한 선택도 추정 및 근사 질의 결과 제시 방법은 객체의 분포가 극단적으로 일정 지역에 집중된 경우 보다는 도시 환경과 같은 경우 적합하며, 질의 시간이 수시간 이후의 미래 질의인 경우에도 대체적으로 정확도를 유지하는 장점을 보였다.

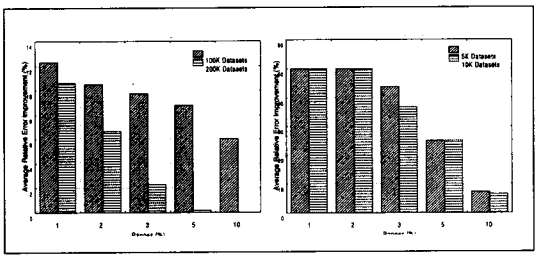


그림 10 데이타 크기에 따른 히스토그램 정확도 개선

6. 결 론

이동 객체의 예상되는 궤적을 다루는 시공간 데이타베이스에서 미래 질의에 대한 선택도 추정을 위한 히스토그램은 객체의 위치 정보가 가지는 특성으로 기존의 방법을 사용하기 곤란하므로 본 논문에서는 질의 술어에 따른 동적 시공간 히스토그램을 생성하는 효율적인 방법을 제안하였다.

본 논문의 기여도는 다음과 같이 요약된다.

- 시공간 데이타 분포를 효율적으로 요약할 수 있는 시놉시스를 이용하여 한계 분포에 의한 시공간 히스토그램 생성 방법: 이동 객체의 궤적이 가지는 시공간 특성을 한계 분포를 이용하여 히스토그램 버킷이 가지는 정보를 엔트로피로 표현하는 과정으로 추상화함으로써 질의에 대한 실시간 히스토그램 생성이 가능

하며, 영역 질의에 대하여 선택도 추정 및 근사 질의 결과를 구하는 효율적인 방법을 제시하였다.

- 데이타베이스의 갱신 내용을 효율적으로 히스토그램에 반영하여 선택도 추정 정확성을 향상: 데이타 요약을 쌍대 공간 정보를 이용함으로써 일정 시간 동안의 객체 위치 정보를 유지할 수 있으며 갱신이 발생하는 경우에도 선형 시간에 이를 시놉시스에 반영함으로써 히스토그램 생성과 선택도 추정 과정 변화없이 데이타 베이스 갱신 내용이 추정 결과에 반영될 수 있도록 하였다.

향후 연구 과제로서 시놉시스를 생성하기 위하여 다양한 공간 분할 방법의 적용을 연구하고 있으며 앞으로 보다 정규화된 시공간 선택도 추정 방법과 이에 대한 정확한 비용 평가 방법을 적용하여 다양한 데이타 집합에 대한 최적의 방법을 선택할 수 있는 방법이 요구된다.

참 고 문 헌

[1] Kollios, G., Gunopulos, D., Tsotras, V., "On Indexing Mobile Objects," Proceedings of PODS, pp. 262-272, 1999.

[2] Park, H.K., Son, J.H., Kim, M.H., "An Efficient Spatiotemporal Indexing Method for Moving Objects in Mobile Communication Environments," The Int. Conf. on MDM, LNCS 2574, pp78-91. 2003.

[3] Wolfson, O., Sistla, P., Chamberlain, S., Yesha, Y., "Updating and Querying Databases that track Mobile Units," J. of Distributed and Parallel Databases, Vol. 7, pp257-287, 1999.

[4] Saltenis, S., Jensen, C., "Indexing of Moving Objects for Location-Based Services," Proceedings of ICDE, pp. 463-472, 2002.

[5] Choi, Y., Chung, C., "Selectivity Estimation for Spatio-Temporal Queries to Moving Objects," Proceedings of SIGMOD Conference, pp. 440-451, 2002.

[6] Thaper, N., Guha, S., Indyk, P., Koudas, N., "Dynamic Multidimensional Histograms," Proceedings of SIGMOD Conference, pp. 427-439, 2002.

[7] Acharya, S., Poosala, V., Ramaswamy, S., "Selectivity Estimation in Spatial Databases," Proceedings of SIGMOD, pp. 13-24, 1999.

[8] Aboulnaga, A., Naughton, J., "Accurate Estimation of the Cost of Spatial Selections," Proceedings of ICDE, pp. 123-134, 2000.

[9] Poosala, V., et al., "Improved Histograms for Selectivity Estimation of Range Predicates," Proceedings of SIGMOD Conference, pp. 294-305, 1996.

[10] Wang, M., Vitter, J., Lim, L., Padmanabhan, S., "Wavelet-Based Cost Estimation for Spatial Queries," The Int. Conf. on SSTD, LNCS 2121, pp. 175-193, 2001.

[11] Chakrabarti, K., et al., "Approximate Query Processing Using Wavelets," Proceedings of the VLDB Conference, pp.111-122, 2000.

[12] Matias, Y., Vitter, J., Wang, M., "Dynamic Maintenance of Wavelet-Based Histograms," Proceedings of 26th VLDB, pp. 101-110, 2000.

[13] Muthukrishnan, S., Poosala, V., Suel, T., "On Rectangular Partitioning in Two Dimensions: Algorithms, Complexity and Applications," The Int. Conf. on DT, LNCS 1540, pp. 236-256, 1998.

[14] Bertimas, D., Tsitsiklis, J., Introduction to Linear Optimization, Athena Scientific, 1997.

[15] Devore, J., Probability and Statistics for Engineering and the Sciences, 5th Ed. Duxbury, Pacific Grove, CA., 2000.

[16] Baeza-Yates, R., Ribeiro-Neto, B., Modern Information Retrieval, Addison-Wesley, 1999.

박 현 규

1987년 육군사관학교 전산학과 학사 1992년 Naval Postgraduate School 전산학과 석사. 1999년~현재 한국과학기술원 전자전산학과 박사과정. 1987년~현재 육군 전산장교. 관심 분야는 시공간 데이타베이스, 위치 기반 시스템, C4I시스템, Pervasive Computing

손 진 현

1996년 서강대학교 전산학과 학사. 1998년 한국과학기술원 전산학과 석사. 2001년 한국과학기술원 전자전산학과 박사 2001년 9월~2002년 8월 한국과학기술원 전자전산학과 박사후 연구원. 2002년 9월~현재 한양대학교 컴퓨터공학과 전임강사. 관심 분야는 데이타베이스, 미들웨어, 워크플로우, 객체지향기술, E-Business

김 명 호

1982년 서울대학교 컴퓨터 공학과 학사 1984년 서울대학교 컴퓨터 공학과 석사 1989년 MICHIGAN 주립대 전산학과 박사. 1989년 MICHIGAN 주립대 연구원 1989년~1993년 한국과학기술원 조교수 1993년~1999년 한국과학기술원 부교수 1999년~현재 한국과학기술원 교수. 1992년~1993년 개방형 컴퓨터 통신 연구회(OSIA). 분산 트랜잭션처리 분과위(TG-TP) 의장. 1996년~1998년 한국정보과학회 DB연구회지 책임편집위원.. 관심 분야는 데이타베이스, 분산트랜잭션, 분산시스템, 워크플로우