시공간 겹침 조인 연산을 위한 선택도 추정 기법
(Selectivity Estimation for Spatio-Temporal a Overlap Join)

이 명 술 † 이 종 연 ††
(Myoung Sul Lee) (Jong Yun Lee)

요 약 시공간 데이타베이스에서 조인 연산은 매우 많은 비용이 소요되며, 시공간 조인 연산의 효율적인 질의 실행 계획을 세우기 위해 조인 연산에 대한 정확한 선택도 추정은 질의처리 성능에 결정적이다. 주어진 두 이산 데이타집합 S_1, S_2의 타임스탬프 t_q에서 시공간 조인 연산은 타임스탬프 t_q에서 서로 교차하는 모든 객체 쌍을 검색하는 것이다. 시공간 조인 연산의 선택도 추정치는 검색된 객체 쌍의 수를 $|S_1 \times S_2|$로 나눈 값이다. 이 논문은 공간 조인 연산의 선택도 추정 기법인 기하 히스토그램 기법을 확장하여 시공간 조인 선택도 추정을 위한 시공간 히스토그램을 제안한다. 균일 데이타 집합과 편중 데이타 집합 모두를 사용하여 제안된 히스토그램 기법으로 시공간 조인 연산의 선택도를 정확하게 추정할 수 있다는 것을 증명하였다. 본 논문의 기여도는 먼저 이산 데이타 집합에 대한 시공간 조인 선택도 추정 연구의 첫 시도를 하였으며 다음으로 이산 객체의 유효시간 동안의 공간 통계정보를 압축하여 히스토그램을 재구축하는 효율적인 유지기법을 제안하였다.

키워드 : 선택도 추정, 시공간 데이타베이스, 서열 데이타, 히스토그램, 겹침 조인

Abstract A spatio-temporal join is an expensive operation that is commonly used in spatio-temporal database systems. In order to generate an efficient query plan for the queries involving spatio-temporal join operations, it is crucial to estimate accurate selectivity for the join operations. Given two dataset S_1, S_2 of discrete data and a timestamp t_q, a spatio-temporal join retrieves all pairs of objects that are intersected each other at t_q. The selectivity of the join operation equals the number of retrieved pairs divided by the cardinality of the Cartesian product $S_1 \times S_2$. In this paper, we propose a spatio-temporal histogram to estimate selectivity of spatio-temporal join by extending existing geometric histogram. By using a wide spectrum of both uniform dataset and skewed dataset, it is shown that our proposed method, called Spatio-Temporal Histogram, can accurately estimate the selectivity of spatio-temporal join. Our contributions can be summarized as follows: First, the selectivity estimation of spatio-temporal join for discrete data has been first attempted. Second, we propose an efficient maintenance method that reconstructs histograms using compression of spatial statistical information during the lifespan of discrete data.

Key words : Selectivity Estimation, Spatio-temporal Databases, Sequence Data, Histogram, Overlap Join

· 이 논문은 2007년도 충북대학교 학술지원사업의 연구비 지원에 의하여 연구되었음

† 학생회원 : 충북대학교 컴퓨터교육과
lee8722@hanmail.net
†† 종신회원 : 충북대학교 컴퓨터교육과 교수
jongyun@chungbuk.ac.kr
(Corresponding author임)

논문접수 : 2007년 8월 3일
심사완료 : 2008년 1월 11일

정보과학회논문지: 데이타베이스 제35권 제1호(2008.2)

1. 서 론

객체 정보가 시간에 따라 계속 변화하는 시공간 데이타베이스에서 조인 연산은 공간적, 시간적 비용이 많이 소요되는 연산이다. 따라서 효율적인 질의 계획을 세우기 위해서는 선택도 추정 연구가 매우 중요하다. 그러나 공간 데이타베이스 분야에서는 모든 연산에 대한 연구가 활발하게 이루어지고 있지만 시공간 데이타베이스 분야에서는 선택(select) 연산에 대한 연구가 대부분이다. 최근에 공간 객체들의 이력 정보를 효과적으로 다룰 수 있는 기술에 대한 연구가 활발하게 이루어지고 있다[1]. 이

러한 연구 결과는 지리 정보 시스템, 토지 관리 시스템, 도시 계획 시스템 등의 시간에 따라 변화하는 공간 정보들을 처리하는 시스템의 구현에 활용될 수 있다.

공간 데이타베이스 분야에서 조인 연산은 객체의 유형에 따라 거리 조인(distance join), 교차 조인(intersect join), 포함 조인(containment join) 및 겹침 조인(overlap join) 등으로 분류된다. 사각형 객체에 대한 공간 조인 연산은 일반적으로 겹침 조인을 의미하며, 본 논문에서도 시공간 조인 연산을 공간 영역에 대한 겹침 조인 연산으로 간주한다. 시공간 데이타베이스에서 두 데이타 집합 S_1, S_2가 주어질 때, 타임스탬프 t_q에서의 시공간 조인 연산은 $\{(o_1, o_2) | o_1 \cap o_2 \neq \varnothing, o_1 \in S_{1_{t=t_q}}, o_2 \in S_{2_{t=t_q}}\}$을 만족하는 객체 쌍의 검색을 의미하며, 시공간 조인 연산의 선택도는 검색된 객체 쌍의 추정치를 S_1과 S_2의 카티션 프로덕트 $|S_1 \times S_2|$로 나눈 값을 의미한다. 단, $S_{1_{t=t_q}}$은 타임스탬프 t_q에서의 S_1이며, $S_{2_{t=t_q}}$는 타임스탬프 t_q에서의 S_2이다. 시공간 데이타베이스는 시간에 따라 공간 정보가 변하는 이산 데이타(discrete data)와 객체 분포가 변하는 즉 객체가 이동하는 시계열 데이타(time-series data)로 분류할 수 있다. 본 논문에서는 이산 데이타에 대한 시공간 조인 연산의 선택도 추정 기법을 제안하고자 한다. 이산 데이타에 대한 질의의 예를 들면, "2000년도에 미국 로스앤젤스(LA) 지역에 건설된 주택을 검색하라."를 들 수 있다.

시공간 데이타베이스에 대한 선택도 추정 연구가 활발하게 이루어지고 있지만 대부분의 연구가 점 질의와 윈도우 질의 같은 선택 연산에 집중되고 있고 조인 연산에 대한 연구는 거의 이루어지지 않고 있다[1-4]. 시공간 데이타베이스에서 조인 연산은 매우 복잡하며, 시간과 공간 모두 고비용 연산이므로 효율적인 질의 실행 계획을 수립하기 위해서는 선택도 추정 기법에 대한 연구가 필요하다.

본 논문에서는 시공간 데이타베이스의 이산 데이타에 대한 조인 연산의 선택도 추정을 위하여 시공간 히스토그램 기법을 제안한다. 시공간 히스토그램은 시간에 대한 점 질의를 갖는 사각형 공간 객체에 대한 조인 연산의 선택도 추정에 사용된다. 시공간 히스토그램은 공간 데이타베이스에서 효율성이 증명된 기하 히스토그램 기법과 시간 영역에 대한 시공간 최소 편중 히스토그램 구조를 조합한 3차원 히스토그램을 의미한다[4,5]. 또한 본 논문에서는 시간이 경과함에 따라 히스토그램을 효율적으로 관리할 수 있는 기법을 제안하며, 실험을 통해 제안한 시공간 히스토그램의 성능을 평가한다. 제안한 시공간 히스토그램의 학문적인 기여도는 다음의 세 가지로 요약할 수 있다. 첫째, 이산 데이타를 갖는 시공간 데이타베이스에서 조인 연산의 선택도 추정에 대한 연구를 활성화시키며, 둘째, 과거부터 현재까지의 이산 데이타에 대한 조인 연산의 선택도 추정이 가능하며, 셋째, 히스토그램의 구축과 유지에 필요한 관리 기법 등을 들 수 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 지금까지 연구된 공간 데이타베이스와 시공간 데이타베이스에서의 선택 연산과 조인 연산의 선택도 추정기법을 검토한다. 제 3장에서는 시공간 히스토그램을 구축하고 관리하는 방법에 대하여 설명하며, 시공간 히스토그램을 이용하여 조인 연산의 선택도를 추정하는 기법에 대하여 기술한다. 제 4장에서는 실험을 통해 제안한 시공간 히스토그램의 성능을 평가하며, 마지막으로 제 5장에서는 본 논문의 결론 및 향후 연구과제에 대하여 기술한다.

2. 관련 연구

이 장에서는 공간 조인 연산과 시공간 선택 연산의 선택도 추정에 대한 연구 배경을 요약한다. 먼저 제 2.1절에서는 공간 조인 연산의 선택도 추정에 관한 기존 연구 결과를 요약하고, 제 2.2절에서는 기하 히스토그램 기법을 이용한 공간 조인 연산의 선택도 추정 기법을 검토하고, 제 2.3절에서는 시공간 데이타베이스에서 시공간 최소 편중 히스토그램(T-Minskew) 기법을 이용한 이산 데이타의 선택 연산에 대한 선택도 추정 기법을 검토한다.

2.1 공간조인 연산의 선택도 추정

공간 데이타베이스(spatial databases)에서 조인 연산의 선택도 추정 연구는 공간 객체의 기하 유형에 따라 크게 점(points), 선(lines) 및 영역(area) 객체 유형에 대한 연구로 나누어진다. 단, 선과 영역 객체는 최소경계 사각형(MBR: Minimum Bounding Rectangle)에 의해 표현될 수 있으므로 사각형 유형으로 취급한다. 점 객체에 대한 공간 조인의 선택도 추정 연구는 프랙탈(fractal) 차원을 이용하는 연구와 두 개의 다차원 점집합 사이의 상호 거리 분포가 멱함수를 따른다는 특성을 이용하는 연구를 들 수 있다[6,7]. 이러한 연구들은 공간 객체의 분포에 의존하며 사각형 객체로 쉽게 확장할 수 없다는 문제점이 있다. 사각형 객체에 대한 공간 조인 연산의 선택도 추정에 대한 연구로 R-트리를 이용한 추정 모델과 조인 연산에 대한 비용 모델을 들 수 있다[8,9].

또한 공간 조인 연산의 선택도 추정 연구에는 통계 정보를 저장하여 이용하는 기법이 있다. 각 공간 데이타 집합에 하나 이상의 통계 정보를 사용하는 매개변수 기법[6,7], 각 공간 데이타 집합에 하나의 히스토그램을 사용하는 히스토그램 기법[5,10,11] 및 이 두 기법을 혼용하는 하이브리드(Hybrid) 기법[12]이 있다. 매개변수 기

법은 객체 분포에 대한 가정을 전제로 하기 때문에 객체 분포에 의존하는 단점을 갖지만 실제 데이타 집합의 특성에 적합한 객체 분포를 사용할 경우에 매우 효율적이다. 선택도 추정 연구 분야에서 일반적으로 사용되는 방법은 히스토그램 기법이며, 히스토그램 기법은 셀 기반의 작은 격자 구조에 통계 정보를 저장하여 선택도를 추정한다. 히스토그램 기법은 객체 분포가 균일해야 한다는 가정을 전제로 하기 때문에 편중 객체 분포에 대해서는 추정 오류율을 줄이는 것이 필요하며, 기하 히스토그램 기법은 이러한 문제를 해결할 수 있는 효율적인 기법으로 알려져 있다[5].

최근 시계열 데이타에 대한 조인 연산의 선택도 추정 연구가 이루어지고 있지만 이산 데이타에 대한 조인 연산의 선택도 추정에 대한 연구는 아직 활성화되지 않고 있다[13]. 조인 연산의 선택도 추정에 대한 연구도 선택 연산의 연구에서와 같이 공간 조인 연산을 기반으로 시공간 조인 연산으로 확장시키는 접근법이 필요하다. 최근에 발표된 시공간 데이타베이스의 선택도 추정 연구에서는 과거 및 현재 시점에서의 이산 데이타에 대한 선택 연산의 선택도 추정 기법을 처음으로 제안하였다[4].

2.2 기하 히스토그램

공간 데이타베이스에서 조인 연산의 선택도 추정을 위한 기하 히스토그램 기법은 원시 객체 집합으로부터 보조 데이타 구조인 히스토그램 파일을 미리 구성한다[5]. 기하 히스토그램 기법은 전체 공간 영역에 객체가 균일하게 분포되었다고 가정한다. 그리고 초기 데이타베이스에서 두 데이타 집합에 대한 조인 연산의 선택도 추정을 단순화하기 위해 두 데이타 집합 중 하나의 데이타 집합을 질의 윈도우의 집합으로 간주하여 윈도우 질의의 선택도의 합이 공간 조인 선택도 추정치와 동일하다는 매개변수 히스토그램 기법을 기초로 한다[14-16].

기하 히스토그램 기법의 기본 개념은 "두 사각형이 교차하여 생기는 교차 영역은 4개의 점을 갖는 사각형이 된다."는 도형의 기본 특징에서 출발한다. 두 사각형의 MBR이 교차하기 위해서는 한 MBR의 꼭지점이 다른 MBR 내부에 위치하거나 한 MBR의 수평선이 다른 MBR의 수직선과 교차해야 한다. 두 사각형이 교차하여 만들어진 교차 영역은 사각형이기 때문에 교차 영역의 꼭지점의 수는 4개이다. 기하 히스토그램 기법은 두 데이타 집합 사이에서 존재하는 교차 영역의 꼭지점의 수를 추정하여 전체 꼭지점의 수를 4로 나누어 전체 교차쌍의 수를 추정하는 기법이다.

따라서 각 버킷을 통과하는 객체의 수직선 수 $V_k(i,j)$와 수평선의 수 $H_k(i,j)$, 각 버킷과 교차하는 객체의 수 $I_k(i,j)$, 버킷 내부에 들어가 있는 객체의 꼭지점 수 $C_k(i,j)$를 히스토그램 파일에 저장하여야 한다. 이들 4

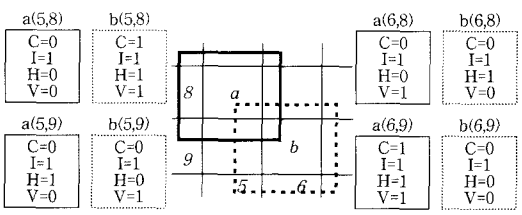


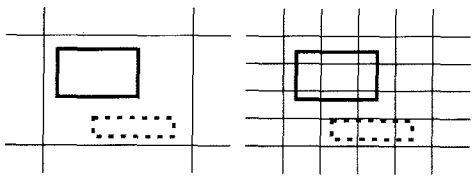
그림 1 기하 히스토그램의 예[5]

가지 정보를 이용하여 데이타 집합 DS_1과 DS_2사이의 교차점을 추정하기 위한 공식은 식 (1)과 같다. 기하 히스토그램의 예는 그림 1과 같다. 그림 1은 데이타 집합 DS_1의 한 객체 a와 데이타 집합 DS_2의 한 객체 b를 전체 객체 공간을 버킷으로 분할한 히스토그램 위에 나타낸 그림이다. 또한 전체 교차점의 수를 추정하기 위해 각 버킷에 저장하여야 할 공간 통계정보를 나타낸다.

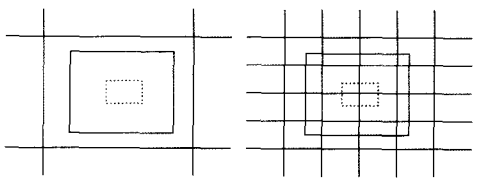
$$N_{a-b} = \sum (C_a(i,j) \times I_b(i,j) + I_a(i,j) \times C_b(i,j) + V_a(i,j) \times H_b(i,j) + H_a(i,j) \times V_b(i,j)) \quad (1)$$

식 (1)의 처음 두 항은 한 MBR의 꼭지점이 다른 MBR 내부에 위치하여 만들어진 교차점을 계산하며, 마지막 두 항은 서로 교차하는 두 MBR에 의하여 만들어진 교차점을 계산한다.

기하 히스토그램 기법은 주어진 버킷 안에 데이타 집합의 모든 MBR의 꼭지점이 이 버킷과 교차하는 다른 데이타 집합의 모든 MBR 내부에 위치해야 하며, 이 버킷과 교차하는 한 데이타 집합 MBR의 모든 수평선이 이 버킷과 교차하는 다른 데이타 집합 MBR의 모든 수직선과 교차하여야 한다는 기본 가정이 필요하다. 그림 2는 기하 히스토그램의 부정확성을 나타내는 예이다. 기하 히스토그램의 부정확성을 감소시키기 위하여 격자의



(a) 카운트 오류



(b) 중복 카운트

그림 2 기하 히스토그램의 부정확성[5]

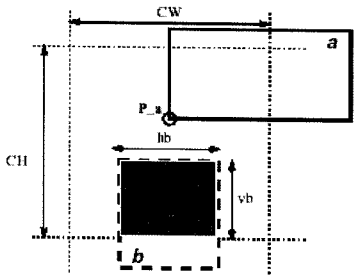


그림 3 정확한 교차점 추정을 위한 기하 히스토그램 수정[5]

표 1 향상된 기하 히스토그램의 변수

변수	의 미
$C_k(i,j)$	셀(i, j) 안의 꼭지점 수
$O_k(i,j)$	셀(i, j)과 교차하는 객체의 교차영역과 셀 크기의 비율 합
$H_k(i,j)$	셀(i, j)과 교차하는 객체의 수평선과 셀의 수평선 길이의 비율 합
$V_k(i,j)$	셀(i, j)과 교차하는 객체의 수직선과 셀의 수직선 길이의 비율 합

해상도를 높여서 해결할 수 있지만 저장 공간과 처리 시간 측면에서 많은 비용이 발생하여 비실용적이 된다.

그림 3은 기존의 기하 히스토그램의 부정확성을 줄이기 위해 확률적으로 교차점의 수를 좀 더 정확히 추정하는 기법을 나타낸다. 그림 3에서 b안에 P_a가 있을 확률은 색칠된 영역 I_b 안에 꼭지점 P_a가 위치할 확률과 같으며, 이를 균일한 분포라 가정한 경우, 셀의 넓이와 I_b의 넓이의 비율이다. 따라서 셀 내부에서 DS_1의 객체의 꼭지점이 N개 있다면 I_b와 교차하는 꼭지점의 개수는 $(N\times\frac{hb\times vb}{CW\times CH})$개라고 할 수 있다. 따라서, 셀$(i, j)$에서 생성되는 전체 교차점의 개수는 $O_2(i,j)\times C_1(i,j)+O_1(i,j)\times C_2(i,j)$개이다. 마찬가지로 넓이가 $CW\times CH$인 2차원 공간 안에서 길이가 v인 수직선과 길이가 h인 수평선이 교차 확률은 $\frac{h\times v}{CW\times CH}$이다. 셀$(i, j)$에서 DS_1의 모든 수직선과 DS_2의 모든 수평선의 교차 수는 $H_2(i,j)\times V_1(i,j)$이고 셀(i, j)에서 DS_1의 모든 수평선과 DS_2의 모든 수직선의 교차 수는 $H_1(i,j)\times V_2(i,j)$이다. 결론적으로 전체 교차점의 수 IP는 식 (2-2)과 같으며 조인 연산의 결과 추정치는 $\frac{IP}{4}$이다.

$$IP=\sum(C_1(i,j)\times O_2(i,j)+C_2(i,j)\times O_1(i,j)+H_1(i,j)\times V_2(i,j)+H_2(i,j)\times V_1(i,j)) \quad (2)$$

2.3 시공간 최소 편중 히스토그램

시공간 데이타베이스의 이산 데이타에 대한 선택 연산의 선택도 추정 기법인 시공간 최소 편중 히스토그램은 공간 데이타베이스 선택도 추정기법인 최소 편중(Minskew) 히스토그램을 확장하였으며, 최소 편중 히스토그램과 마찬가지로 객체의 분포가 균일하지 못할 경우에는 히스토그램의 재구축을 통하여 일정 수준의 오류율을 유지하도록 하였다[17].

객체의 삽입, 삭제 연산으로 인하여 히스토그램을 갱신해야 할 경우 히스토그램 재구축 조건을 충족시키지 못한다면 해당 객체에 대한 객체 변화를 해당 버킷 내에 변화 정보를 시간 속성과 함께 버킷에 저장한다. 히스토그램 재구축은 한 히스토그램이 유지되는 동안 전체 객체 수에 대한 객체 변화 횟수가 주어진 임계치를 넘어설 경우에 발생한다. 그리고 그림 4와 같이 새롭게 구축된 히스토그램 이전에 존재하던 히스토그램 정보는 유효 시간에 재구축 시점을 할당하여 이력 정보로서 저장된다.

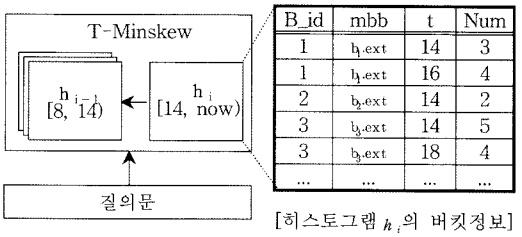


그림 4 시공간 최소 편중 히스토그램의 구조

시공간 이력 질의에 대한 선택도 추정을 위해서는 우선 질의 타임스탬프와 겹치는 히스토그램을 검색하고, 히스토그램 내에서 질의 공간 영역과 겹치는 버킷들을 검색한다. 질의 공간 영역과 겹치는 모든 버킷에 대하여 질의와 겹치는 버킷 넓이와 버킷 전체 넓이의 비율을 구하여 버킷에 포함된 객체 수를 곱하면 질의와 겹쳐지는 버킷 내의 객체 수를 추정할 수 있다. 마지막으로, 각 버킷에서 추정된 객체 수를 더하여 전체 선택도를 추정한다. j번째 버킷을 B_j, 그 버킷 내의 객체 수를 $B_j.num$, j번째 버킷의 MBR을 $B_j.MBR$, j번째 버킷의 선택도를 Sel_j라 할 때, 식 (3)과 식 (4)는 선택도를 계산하는 식이다.

$$Sel_j = B_j.num\times Overlap(B_j.MBR)/area(B_j.MBR) \quad (3)$$

$$Sel=\sum_{j=0}^{k} Sel_j \quad (4)$$

시공간 최소 편중 히스토그램의 장점은 히스토그램의 재구축 시점에서 기존의 정보를 이력 정보로 사용하여

과거 시점에 대한 선택도 추정도 가능하다는 점이다. 그러나 선택도 추정을 위한 질의가 특정 시점에서 공간 영역을 가지는 질의로 제한되기 때문에 시간적 범위를 갖는 질의가 불가능하다는 단점을 가지고 있다. 또한, 히스토그램 특성상 객체의 차원이 높아질수록 히스토그램의 저장 용량 면에서 과중한 부담과 높은 오류율이 발생한다는 단점을 가지고 있다.

3. 시공간 히스토그램 구축 및 선택도 추정

3.1 시공간 히스토그램의 구조

시계열 데이타 집합의 객체는 시간 경과에 따라 위치가 계속 변경되는 반면에 이산 데이타 집합의 객체는 일정한 시간 주기로 객체가 생성 또는 소멸된다. 특히 이산 데이타의 경우에는 비교적 객체의 생명 주기가 길다는 특성을 가지고 있다. 예를 들면 토지관리 시스템의 경우에 토지의 용도가 한번 정해지면 그 용도가 쉽사리 변경되지 않기 때문에 객체의 유효 시간이 매우 길다. 시공간 데이타베이스에서 조인 연산의 선택도 추정에 히스토그램 기법을 이용할 경우 각 타임스탬프에 해당하는 객체의 공간정보를 저장하는데 많은 비용이 소요된다는 문제가 발생한다. 이 장에서는 시간 경과에 따라 객체의 공간정보 변경이 비교적 적게 일어나는 이산 데이타 집합의 특성을 이용하여 객체 정보를 포함하고 있는 히스토그램의 셀들을 시간 축으로 압축하여 공간 및 시간 비용을 감소시킬 수 있는 시공간 히스토그램 기법을 제안한다.

시공간 히스토그램은 공간 조인 연산의 선택도 추정에 사용된 기하 히스토그램에 시간 축을 추가한 3차원 히스토그램이다. 특정 질의 타임스탬프 t_q에서 두 데이타 집합에 대한 시공간 조인 연산의 선택도 추정은 우선 타임스탬프 t_q에서 겹치는 두 데이타 집합의 히스토그램을 검색한 후 기하 히스토그램 알고리즘을 이용하여 선택도 추정치를 계산한다.

히스토그램의 재구축은 객체가 생성 또는 소멸되는 시점에서 발생되며, 객체의 공간 정보를 이용하여 객체와 겹치는 히스토그램의 셀들을 검색하여 공간 통계 정보를 갱신한다. 객체가 갱신되면 이전 히스토그램의 유효시간의 끝 시간이 현재 타임스탬프로 변경되며, 새로 생성된 히스토그램의 유효시간의 시작 시간도 현재 타임스탬프가 된다.

본 논문에서 사용하는 시공간 조인 질의는 특정 질의 타임스탬프 t_q에서의 질의이며, 공간 영역의 전체 도메인 공간은 $2^n \times 2^n$(단, n은 셀 레벨)개의 셀로 분할되고 시간 영역은 $(r+1)$개의 셀로 나누어지며 r은 재구축 횟수를 의미한다. 따라서 시공간 히스토그램에 포함할 수 있는 전체 셀의 수는 $2^n \times 2^n \times (r+1)$이 된다. 각 셀에는 두 데이타 집합의 조인 연산 시 결과 쌍의 수를 추정하는데 사용되는 객체의 공간 정보인 c, w, h, a 값을 저장한다. 표 2는 본 논문에서 자주 사용하는 변수들의 종류와 각각의 의미를 나타낸다.

표 2 변수 정의

변수	설 명
N_{join}	시공간 조인 연산의 결과 쌍의 수
t_q	질의 시간의 타임스탬프
H_{t_i}	유효시간의 시작 타임스탬프가 t_i인 히스토그램
ls	유효시간의 시작 타임스탬프
le	유효시간의 끝 타임스탬프
$X_{\min}$	객체의 X축 최소 좌표
$Y_{\min}$	객체의 Y축 최소 좌표
$X_{\max}$	객체의 X축 최대 좌표
$Y_{\max}$	객체의 Y축 최대 좌표
c	셀 안에 있는 객체의 꼭지점 수
w	각 셀과 겹치는 객체의 가로 길이와 셀 크기에 대한 비율의 합
h	각 셀과 겹치는 객체의 세로 길이와 셀 크기에 대한 비율의 합
a	각 셀과 겹치는 객체의 넓이와 셀 넓이에 대한 비율의 합

3.2 셀 초기화

타임스탬프 t_0에서 $2^n \times 2^n$개의 전체 초기화된 셀들을 생성하여 히스토그램 H_0를 생성하며, 히스토그램의 유효시간을 [0, now)로 초기화한다. 객체의 공간 비용을 감소시키기 위해 셀에는 공간 객체에 대한 통계 정보가 저장된 주소를 가리키는 포인터가 저장된다. 변수 c, w, h, a 값을 초기화하며 변수들이 저장된 주소를 각 셀에 할당한다. 타임스탬프 t_0에서 생성된 공간 객체와 교차하는 셀들을 검색하여 객체의 공간 통계 정보를 갱신, 저장하고 주소를 추가되는 공간 객체와 교차하는 셀에 할당한다.

타임스탬프 t_0이후 가장 최근에 생성 또는 소멸된 객체를 검색하여 히스토그램을 차례로 구축한다. 예를 들어 타임스탬프 t_0 이후에 객체가 생성 또는 삭제되는 사건이 타임스탬프 t_i에서 발생하였다면 히스토그램 H_i를 생성하여 유효시간을 [t_i, now)로 할당한다. 그리고 타임스탬프 t_i에서 생성 또는 소멸된 객체와 교차하는 셀들을 이전 히스토그램에서 검색하여 객체의 공간 통계정보를 계산, 저장하고 주소를 히스토그램 H_i의 셀들에 할당한다. 그리고 H_i의 나머지 셀들의 주소는 이전 히스토그램의 셀이 가지고 있는 주소로 저장된다.

3.3 교차점의 개수 계산

기하 히스토그램 기법에 의하면 특정 타임스탬프 t_q에서 시공간 조인 연산의 결과 쌍의 추정치는 식 (5)와 같다.

$$IP=\sum_{i,j}(c_1(i,j)\times a_2(i,j)+c_2(i,j)\times a_1(i,j)+h_1(i,j)\times w_2(i,j)+h_2(i,j)\times w_1(i,j))N_{join}=\frac{IP}{4} \quad (5)$$

표 3은 시공간 히스토그램을 이용한 시공간 조인 연산의 선택도 추정을 설명하기 위하여 데이타 집합 A, B 각각에 대한 4개의 객체를 나타낸 표이다. 각 셀은 시공간 객체의 공간 정보인 MBR $\{X_{\min}, Y_{\min}, X_{\max}, Y_{\max}\}$과 시간 정보인 유효시간(lifespan)을 나타낸다. 두 데이타 집합의 조인 연산을 위하여 표 3의 각 객체를 2차원 격자 공간에 표현하면 그림 5와 같다. 그림 5에서 한 셀의 크기는 1이고 가로는 객체의 x좌표, 세로는 객체의 y좌표에 해당한다. 그림 5에서 실제 조인한 결과 쌍은 $(a_1,b_2),(a_2,b_1),(a_2,b_2),(a_3,b_2)$으로 4쌍이다. 그림 6과 그림 7은 두 데이타 집합 A, B에 대하여 각 객체의 공간 정보 c, w, h, a 값을 계산하여 각 셀에 저장한 시공간 히스토그램 H_{12}^A 및 H_{12}^B를 나타낸다. 히스토그램 H_{12}^A에서 셀(x_1,y_1)의 c=1, a=0.25(0.5×0.5), w=0.5, h=0.5이다.

표 3 두 개의 데이타 집합 A, B에 대한 시공간 객체의 표현 사례

Dataset	ID	$X_{\min}$	$X_{\max}$	$Y_{\min}$	$Y_{\max}$	유효시간
A	a_1	0.5	1.5	2.5	3.5	[0, now)
	a_2	1.66	3.33	1.66	3.33	[3, now)
	a_3	0.5	1.5	0.5	1.5	[8, now)
	a_4	1.66	2.33	0.33	0.66	[12, now)
B	b_1	2.5	3.5	2.5	3.66	[0, now)
	b_2	1.33	2.33	1.33	2.66	[3, now)
	b_3	2.5	3.5	0.5	1.5	[8, now)
	b_4	0.33	1.5	1.66	2.33	[12, now)

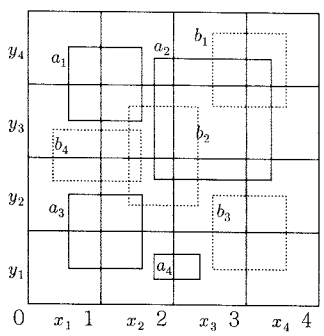


그림 5 타임스탬프 t_q에서 두 개의 데이타 집합 (데이타 집합 A : 실선, 데이타 집합 B : 점선)

y_4	c=1 a=0.25 w=0.5 h=0.5	c=2 a=0.36 w=0.83 h=0.83	c=0 a=0.33 w=1 h=0	c=1 a=0.109 w=0.33 h=0.33
y_3	c=1 a=0.25 w=0.5 h=0.5	c=1 a=0.58 w=0.5 h=1.5	c=0 a=1 w=0 h=0	c=0 a=0.33 w=0 h=1
y_2	c=1 a=0.25 w=0.5 h=0.5	c=2 a=0.36 w=0.83 h=0.83	c=0 a=0.33 w=1 h=0	c=1 a=0.109 w=0.33 h=0.33
y_1	c=1 a=0.25 w=0.5 h=0.5	c=3 a=0.359 w=0.83 h=0.83	c=2 a=0.109 w=0.33 h=0.33	c=0 a=0 w=0 h=0
	x_1	x_2	x_3	x_4

그림 6 데이타 집합 A의 H_{12}^A

y_4	c=0 a=0 w=0 h=0	c=0 a=0 w=0 h=0	c=1 a=0.33 w=0.5 h=0.66	c=1 a=0.33 w=0.5 h=0.66
y_3	c=1 a=0.218 w=0.66 h=0.33	c=2 a=0.601 w=1.16 h=0.99	c=2 a=0.468 w=0.83 h=0.16	c=1 a=0.25 w=0.5 h=0.5
y_2	c=1 a=0.218 w=0.66 h=0.33	c=2 a=0.601 w=1.16 h=0.99	c=2 a=0.468 w=0.83 h=0.16	c=1 a=0.25 w=0.5 h=0.5
y_1	c=0 a=0 w=0 h=0	c=0 a=0 w=0 h=0	c=1 a=0.25 w=0.5 h=0.5	c=1 a=0.25 w=0.5 h=0.5
	x_1	x_2	x_3	x_4

그림 7 데이타 집합 B의 H_{12}^B

Algorithm for computing the number of intersection point
Begin
1: Find a Histogram H_i^A, H_i^B
2: that satisfy $H_i^A.ls \leq t_q < H_i^A.le$ and $H_i^B.ls \leq t_q < H_i^B.le$
3: where $0 \leq i \leq$ number of histogram;
4: For j=0 to the number of cells
5: retrieve the information of cells;
6: $Ip_j = c_j^{A*} a_j^B + c_j^{B*} a_j^A + w_j^{A*} h_j^B + w_j^{B*} h_j^A$;
7: $Ip = Ip + Ip_j$;
8: End for
End

그림 8 교차점 개수의 계산 알고리즘

히스토그램의 각 셀에는 셀과 교차하는 꼭지점 개수의 합(c), 셀과 겹치는 객체의 넓이와 셀 넓이간의 비율의 합(a), 셀과 겹치는 객체의 가로 길이와 셀 크기간의 비율의 합(w) 및 셀과 겹치는 객체의 세로 길이와 셀 크기간의 비율의 합(h)을 계산하여 저장한다. 그림 8은 교차점의 개수 I_p를 구하는 알고리즘이다. 먼저 질의 타

임스탬프 t_q에 해당하는 히스토그램 H_i^A, H_i^B을 히스토그램 유효시간의 시작 시간과 끝 시간을 질의 타임스탬프 t_q와 비교하여 두 데이타 집합의 시공간 히스토그램에서 검색한다. 다음으로 각 셀에 할당된 주소 정보를 이용하여 각 셀과 교차하는 객체의 공간 통계 정보를 가져온 후에 교차점의 개수 I_p를 계산한다. 그림 6과 그림 7의 히스토그램에 알고리즘을 적용했을 경우 교차점의 개수는 I_p는 17.469가 된다. 결론적으로, 객체의 교차 쌍의 개수는 4.367이 되며 그림 5에서와 같이 실제 두 데이타 집합 A, B의 객체들에 대한 교차 쌍의 개수 4와 거의 일치하게 된다.

3.4 셀 정보 검색

그림 9는 표 3의 데이타 집합 A에 포함된 각 객체를 유효시간에 해당하는 3차원 격자로 표현한 그림에서 시간 축과 x축만을 나타낸 그림이다. 그림 10은 객체가 생성 또는 소멸되면서 히스토그램이 재구축되었을 경우 셀 정보가 변경된 히스토그램을 나타낸다. 히스토그램의 셀들 중에 변화가 일어나지 않는 셀에는 객체의 공간 통계정보가 기록되지 않는다.

시공간 히스토그램의 중요한 특징으로는 그림 9에서와 같이 객체가 생성 또는 소멸되는 시점에서 히스토그램이 재구축되기 때문에 히스토그램 유효시간의 시작 또는 끝 시간이 객체 유효시간의 시작 또는 끝 시간과 일치한다는 점이다. 따라서 객체의 가로 경계선이 히스토그램의 가로 경계선과 정확히 일치된다. 히스토그램의 셀의 경계선에 객체가 일치되기 때문에 히스토그램의 유효시간 안에 있는 질의 타임스탬프에 대한 선택도 추정의 정확성이 보장된다. 그림 9에서 히스토그램 H_3의 유효시간이 [3, 8)이라면 [3, 8) 시간 내에서는 객체가 갱신되지 않았다는 것을 의미하며 타임스탬프 3에서 타임스탬프 8까지 셀 정보는 모두 동일하게 된다. 중간 타임스탬프 5에서의 질의가 이루어졌을 경우에도 히스토그램 H_3를 이용하면 정확한 선택도 추정이 가능하다.

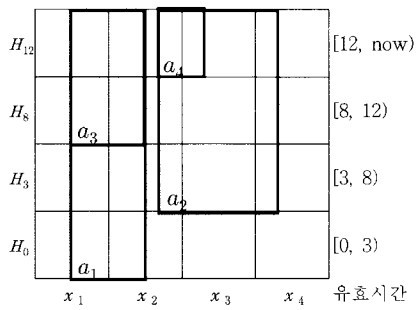


그림 9 유효시간을 볼 수 있는 객체의 분포도 (x축, 시간축만 표현)

H_{12}	H_8	c, a, w, h	c, a, w, h	H_3
H_8	c, a, w, h	c, a, w, h	H_3	H_3
H_3	H_0	c, a, w, h	c, a, w, h	c, a, w, h
H_0	c, a, w, h	c, a, w, h	c, a, w, h	c, a, w, h
	x_1	x_2	x_3	x_4

그림 10 셀을 압축한 시공간 히스토그램

그림 9의 객체 a_1이 히스토그램 H_0에 유지되는 중에 타임스탬프 3에 새로운 객체 a_2가 생성되어 히스토그램의 재구축이 발생하면 새로운 히스토그램 H_3가 생성되면서 H_3의 셀 x_2, x_3, x_4는 객체가 추가되어 셀 정보가 변화된다. 하지만 x_1은 객체 a_2의 MBR과 교차하지 않는 셀이므로 셀 정보가 변화하지 않는다. 따라서 그림 10과 같이 히스토그램 H_3의 셀 x_1은 새로운 객체의 공간 통계정보가 아닌 히스토그램 H_0의 셀 정보를 소유하게 된다. 시공간 데이타베이스의 조인 연산 선택도 추정에서 발생하는 과다한 시공간 비용 문제는 객체의 유효시간이 비교적 긴 이산 데이타의 특성과 객체 압축 방법을 이용하여 해결할 수 있다.

그림 11은 두 데이타 집합의 객체의 교차 쌍의 수를 계산하기 위하여 히스토그램의 셀 정보를 검색하는 알고리즘이다. 먼저 질의 타임스탬프와 히스토그램의 유효시간을 비교하여 질의 타임스탬프에 해당하는 히스토그램을 검색한다. 다음으로 $cell_j$에 있는 객체의 공간 통계정보를 반환하고 공간 통계정보가 아닌 히스토그램의 주소가 들어 있으면 그 주소에 해당하는 히스토그램의 $cell_j$에 있는 공간 통계정보를 반환한다.

```
Algorithm for retrieving the information of cells
Begin
1:   Find a Histogram H_i
2:      that satisfies H_i.ls ≦ t_q < H_i.le
3:      where 0 ≦ i ≦ number of histogram;
4:   For j=0 to the number of cells
5:      If (cell_j is an address of Histogram)
6:         cell_j = H_{cell_j.ptr}.cell_j;
7:      End if
8:   End for
End
```

그림 11 셀 정보 검색 알고리즘

3.5 시공간 히스토그램의 갱신 및 재구축

시공간 히스토그램의 재구축은 객체가 생성 또는 소멸되는 시점에서 발생한다. 히스토그램의 재구축 횟수가 많이 발생할 수 있지만 생성 또는 소멸되는 객체에 대한 공간 정보만을 셀에 반영하기 때문에 재구축 연산은 매우 간단히 수행된다. 시공간 히스토그램의 재구축 알고리즘은 그림 12와 같다. 객체가 생성 또는 소멸되면 가장 최근의 히스토그램 H_i를 검색하여 히스토그램 H_i의 유효시간 끝 시점을 현재 타임스탬프에 할당하고, 현재 타임스탬프를 유효시간 시작점으로 하는 새로운 히스토그램 H_{now}를 생성한다. 그림 12에서와 같이 히스토그램 H_i의 각 셀의 정보를 검색하여 단계 7~14까지의 과정을 반복 수행한다. 단계 7에서는 셀의 x, y 좌표의 최대값과 최소값이 객체의 MBR 좌표 사이에 있는지를 비교하며, 교차할 경우 단계 8에서 객체와 셀이 교차하는 가로, 세로변의 길이와 넓이를 계산하여 셀 정보에 반영한다. 객체와 셀이 교차하지 않는다면 객체가 변경되어도 셀 정보가 변경되지 않기 때문에 히스토그램 H_i의 셀에 저장된 주소를 새로운 셀에 할당한다.

객체가 생성 또는 소멸되는 경우 객체의 시간 속성 변경으로 인하여 히스토그램의 재구축이 필요하지만 공간 속성만을 변경할 경우 재구축 대신에 해당 히스토그램의 셀 정보만 변경한다. 그림 13은 히스토그램의 공간 통계정보를 갱신하는 알고리즘을 나타낸다. 공간 정보가 변경된 객체의 유효시간 시작점이 i라 할 때, 히스토그램 H_i을 검색, 객체의 과거 공간 정보를 이용하여 c, w, h, a 값을 계산한다. 그리고 히스토그램과 교차하는 셀들을 검색하여 객체의 공간 통계정보 값을 갱신한다. 또한 변경될 객체의 새로운 공간 정보를 이용하여

```
Algorithm for reconstructing the spatio-temporal histogram
Begin
1:   Find a last Histogram H_i
2:      where i = max(i)
3:   ls = current timestamp;
4:   H_i.le = ls;
5:   construct a new histogram H_is;
6:   For j=0 to the number of cells
7:      If (H_i.cell_j overlapped with object in H_i)
8:         compute c, w̄, h̄, ā using x_min, x_max, y_min, y_max;
9:         update H_is.cell_j set c, w̄, h̄, ā;
10:     Else if (cell_j is an address of Histogram)
11:        H_is.cell_j = H_i.cell_j;
12:     Else
13:        H_is.cell_j = an address of Histogram H_i;
14:     End if
15:  End for
End
```

그림 12 시공간 히스토그램 재구축 알고리즘

```
Update the information of cells Algorithm
Begin
     Find a Histogram H_i
1:      that satisfy H_i.ls = o.ls
2:      where 0 ≤ i ≤ number of histogram;
3:   HgNum = o.ls;
4:   For j=0 to the number of cells
5:      If (H_i.cell_j overlapped with old object in H_i)
6:         compute c, w, h, a using old object's MBR;
7:         For (i=HgNum; cell_j is not an address of
8:  Histogram ;i++)
9:            subtract c, w, h, a from H_i.cell_j;
10:        End for
11:     End if
12:     If (H_i.cell_j overlapped with new object in H_i)
13:        compute c, w, h, a using new object's MBR;
14:        For (i=HgNum; cell_j is not an address of
15: Histogram ;i++)
16:           add c, w, h, a from H_i.cell_j;
17:        End for
18:     End if
     End for
End
```

그림 13 시공간 히스토그램의 공간 정보 갱신 알고리즘

셀에 저장될 공간 통계정보 값을 계산하며, 그 값을 히스토그램과 교차하는 셀들의 공간 통계정보에 추가시킨다. 그러나 히스토그램 H_i을 이용하여 다음 히스토그램을 재구축하였기 때문에 히스토그램 H_i 다음에 재구축된 히스토그램을 모두 검색하여 공간 통계정보를 갱신해야 한다. 나중에 구축된 히스토그램에서 같은 위치에 있는 셀 정보 값이 실제 값이 아닌 히스토그램 주소가 나올 때까지 이후에 구축된 히스토그램을 계속 검사하면서 수정한다.

3.6 겹침 조인 연산의 선택도 추정

시간에 대한 점 질의는 조인 연산의 대상이 되는 두 데이타 집합에 대한 질의 타임스탬프 t_q^A, t_q^B가 상이한 질의($t_q^A \neq t_q^B$)와 동일한 질의($t_q^A = t_q^B$)로 나누어진다. 시공간 히스토그램은 각 데이타 집합에 대해 독립적으로 히스토그램이 구축되기 때문에 두 데이타 집합에 대한 질의 타임스탬프가 상이하더라도 선택도 추정이 가능하다.

객체의 생성 또는 소멸 시점에서 시공간 히스토그램을 재구축하기 때문에 각 타임스탬프에 따라 적용되는 히스토그램이 다르다. 따라서 시간에 대한 점 질의를 갖는 시공간 데이타베이스의 조인 연산 선택도 추정을 위해서는 조인 연산의 대상이 되는 두 데이타 집합의 질의 타임스탬프 t_q와 교차하는 히스토그램을 검색하고, 각각의 히스토그램 내에서 공간 영역에 대한 조인 연산의 선택도를 추정한다.

그림 14는 시간에 대한 점 질의를 갖는 시공간 데이타베이스의 조인 연산 선택도 추정을 위해 결과 쌍의

Selectivity estimation algorithm by using spatio-temporal histogram
Begin
1: t_q^A, t_q^B : query timestamp of datasets;
2: H_i^A = retrieve the information of cells at timestamp t_q^A;
3: H_i^B = retrieve the information of cells at timestamp t_q^B;
4: Ip = compute the number of intersection point;
5: $N_{join} = \frac{Ip}{4}$;
End

그림 14 시공간 히스토그램을 이용한 조인 결과 추정 알고리즘

추정치를 구하는 알고리즘을 나타낸다. 단계 1에서는 질의 타임스탬프 t_q에 해당하는 히스토그램 H_i^A, H_j^B을 히스토그램의 유효시간과 질의 타임스탬프 t_q를 비교하여 검색한다. 단계 2~3에서는 각각의 히스토그램 H_i^A, H_j^B을 이용, 선택도를 추정하기 위해 우선 그림 11 알고리즘을 이용하여 셀 정보를 검색한다. 단계 4에서 히스토그램 H_i^A, H_j^B에서 검색한 셀 정보를 이용하여 두 데이타 집합 A와 B의 객체들의 공간 영역의 교차로 발생하는 교차점의 수 I_p를 계산한다. 두 데이타 집합 A와 B의 객체가 서로 한 번씩 교차하여 만들어지는 교차점의 수는 4개이므로 최종 질의 타임스탬프 t_q^A, t_q^B에서 조인 연산의 결과 쌍의 추정치는 $\frac{Ip}{4}$가 된다.

4. 실험 및 결과분석

이 장에서는 시공간 데이타베이스에 대한 조인 연산의 선택도 추정을 위해 본 논문에서 제안한 시공간 히스토그램의 효율성을 실험을 통해 증명하고자 한다. 실험은 인텔 T2500 2.0GHz CPU, 1GB RAM, 100GB HDD의 성능을 갖는 데스크탑 PC의 Windows XP 환경에서 수행하였으며, 시공간 히스토그램의 구현 언어로는 eclipse 3.2의 Java를 사용하였다.

실제 데이타 집합(syntactic dataset)인 GSTD (Generate Spatio-Temporal Dataset)[18]와 실험 데이타 집합(experimental dataset) 모두를 사용하여 제안한 시공간 히스토그램의 정확성을 증명하고자 한다. 먼저 실제 데이타 집합에 대해서는 GSTD를 이용하여 객체의 중심점이 가우스 분포를 이루는 집합과 임의 분포를 이루는 집합을 생성하여 실험하였고 실험 데이타 집합에 대해서는 객체 분포가 균등한 데이타 집합과 편중된 데이타 집합으로 구분하여 실험하였다. 본 실험에서는 한 객체의 크기를 객체의 네 변의 길이의 평균으로 간주한다. 데이타 집합 내의 객체의 크기가 50~300인 객체와 50~500인 객체 모두에 대하여 실험을 수행하여 객체의 크기와 선택도 추정의 정확도간에 상호 의존성이 거의 없다는 것을 증명하고자 한다. 또한 실험에서 사용한 객체의 형태는 사각형 형태를 전제로 하였지만 점, 다각선, 다각형 형태의 모든 객체도 MBR에 의하여 추상화 될 수 있기 때문에 다른 형태의 객체 사용도 가능하다. 실험에서 사용한 실제 실험 데이타와 무작위 실험 데이타는 표 4와 같다. 각각의 데이타 집합의 크기에 따라 25,000개의 객체가 생성되어 소멸되기 때문에 총 50,000회의 연산이 이루어지고 시공간 히스토그램의 구축 또는 재구축이 발생된다. 무작위 실험 데이타의 편중된 객체 분포를 위하여 15,000개 객체의 최소 X, Y 좌표값이 0에서 5000이 되도록 생성하였다. 표 5는 시공간 객체가 생성되는 X, Y축 도메인과 시간 도메인을 나타낸다.

표 4 실험 데이타

종류	객체분포	객체 크기	객체 개수
실험 데이타	균일 분포	50~300	25,000
		50~500	25,000
	편중 분포	50~300	25,000
		50~500	25,000
종류	객체분포	객체 밀도	객체 개수
실제 데이타	가우스 분포	500	25,000
	임의 분포	500	25,000

표 5 시공간 객체가 위치하는 도메인

변수명	도메인
X_DOMAIN	0~12800
Y_DOMAIN	0~12800
T_DOMAIN	0~100

4.1 실험 모델

시공간 히스토그램을 이용하여 질의 결과를 추정한 값 Sel과 실제 질의 결과의 값 Sel'을 이용한 상대 오류율의 계산식은 식 (6)과 같다.

$$E = \frac{|Sel - Sel'|}{Sel'}, \text{ for } Sel' > 0 \quad (6)$$

식 (6)에서 실제 질의 결과 값이 0일 경우 상대 오류율은 $E = |Sel - Sel'|$로 표현한다. 특정 질의에 대해 편중된 결과가 나오는 것을 방지하기 위하여 q_n개의 질의에 대한 선택도 추정치의 상대 오류율을 측정하고, 이들의 평균을 최종 상대 오류율로 정하여 실험 결과가 보다 높은 신뢰도를 갖도록 한다. 따라서 최종 상대 오류율을 계산하는 식은 식 (7)와 같으며 n은 질의 횟수를 의미한다.

$$\overline{E} = \frac{\sum_{i=1}^{n} E_i}{n} \quad (7)$$

본 논문에서는 참고 대상이 되는 실험이 없는 관계로

표 6 실험 종류

실험명칭	종 류	객체분포	객체크기	객체개수
<실험1>	실험데이타	균일분포	50~300	25,000
<실험2>		균일분포	50~500	25,000
<실험3>		편중분포	50~300	25,000
<실험4>		편중분포	50~500	25,000
<실험5>	실제데이타	가우스분포	500(밀도)	25,000

표 6과 같은 특징을 갖는 객체들을 대상으로 실험을 수행하였다.

4.2 균일 분포를 갖는 실험 데이타 집합의 선택도 추정

실험 1과 실험 2에서는 도메인 공간 내에 객체가 균일하게 분포되어 있는 두 데이타 집합에 대하여 시공간 조인 연산의 선택도를 추정하며, 최종 상대 오류율을 계산하여 그 결과를 분석한다. 실험 1에서는 객체의 최대 크기가 300인 25,000개의 객체가 임의적으로 생성 및 소멸되는 데이타 집합에 대하여 격자 셀의 레벨을 조정하여 상대 오류율을 측정하였다. 그리고 각각의 셀 레벨에 대한 최종 상대 오류율은 9개의 타임스탬프에 대한 선택도 추정치의 상대 오류율의 평균값이다. 실험 2에서는 객체의 최대 크기가 500인 25,000개의 객체가 임의적으로 생성 및 소멸되는 데이타 집합에 대해서 실험 1과 동일한 방법으로 실험을 수행한다. 표 7은 실험 데이타의 시공간 객체가 갖는 변수 값을 나타낸다.

균일 객체 분포에 대한 조인 연산의 최종 실험 결과는 그림 15, 그림 16과 같다. 그림 15와 그림 16에서와 같이 상대 오류율이 3.17~5.31%로 나왔으며, 셀 크기에 관계없이 비교적 안정적인 실험결과를 나타내고 있다. 그림 15의 객체의 크기가 500이하일 경우에 상대 오류율이 약간 높은 것은 객체의 크기가 균일한 경우가 객체의 크기에 대한 분산도가 높은 경우보다 정확도가 높다는 것을 의미한다. 또한 균일 객체 분포의 경우 셀 크기에 관계없이 상대 오류율이 비슷하다. 그리고 셀 크기를 객체의 평균 크기 이하로 분할할수록 계산이 중복되므로 과추정(overestimation)이 발생하여 상대 오류율이 증가되지만 무시할 수 있는 오류율이 된다. 객체의 균일성이 보장되는 데이타 집합에 대해서는 셀 크기를 객체의 평균 크기보다 크게 할 경우에 공간비용을 최소화

표 7 균일 분포에서 선택도 추정 실험을 위한 변수와 값

변 수	값
MIN_SIZE	50
MAX_SIZE_1	300
MAX_SIZE_2	500
t_q	10, 20, 30, 40, 50, 60, 70, 80, 90
i	8
SELL_SIZE(i=1~8)	$12800/2^{i-1}$

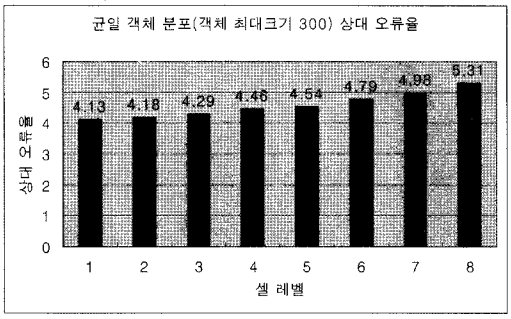


그림 15 <실험 1>의 상대 오류율

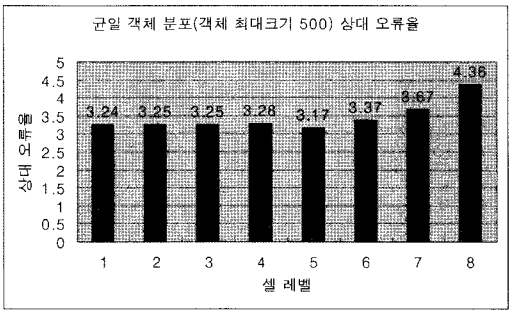


그림 16 <실험 2>의 상대 오류율

할 수 있다.

4.3 편중 분포를 갖는 실험 데이타 집합의 선택도 추정

실험 3과 실험 4에서는 도메인 공간 내에 객체가 특정 좌표를 중심으로 매우 편중되어 있는 두 데이타 집합에 대한 시공간 조인 연산의 선택도를 추정하며, 상대 오류율을 계산하여 그 결과를 분석한다. 실험에서는 실험 1과 실험 2의 경우와 동일한 객체를 사용하였으며, 객체를 편중되게 분포시키기 위해 객체 MBR의 최소 좌표값이 모두 0~5000이 되도록 하였다. 위의 표 8은 실험 데이타의 시공간 객체가 갖는 변수 값을 나타낸다.

편중 객체 분포에서의 조인 연산 선택도 추정치의 최종 실험 결과는 그림 17, 그림 18과 같다. 편중 객체 분포에서의 실험 결과로서 상대 오류율이 셀 크기에 의존

표 8 편중 분포에서 선택도 추정 실험을 위한 변수와 값

변 수	값
MIN_SIZE	50
MAX_SIZE_1	300
MAX_SIZE_2	500
MINX_MIN, MINY_MIN	0
MINX_MAX, MINY_MAX	5000
t_q	10, 20, 30, 40, 50, 60, 70, 80, 90
i	7
SELL_SIZE(i = 1~7)	$12800/2^{i-1}$

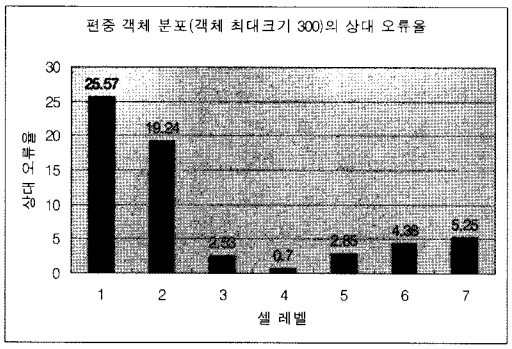


그림 17 실험 3의 상대 오류율

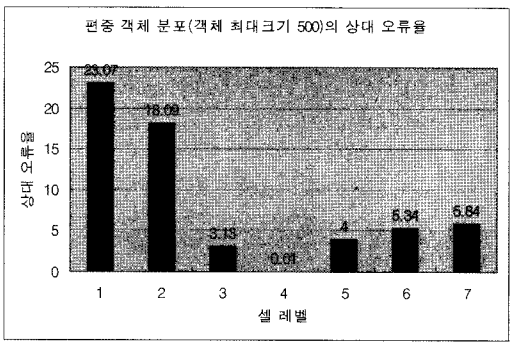


그림 18 실험 4의 상대 오류율

하는 결과를 나타낸다. 셀 크기가 객체의 평균 크기보다 클 경우 선택도 추정치의 상대 오류율이 19.24% 이상이 되므로 정확성이 많이 떨어지는 결과를 나타낸다. 그러나 셀 크기가 객체의 평균 크기보다 작은 경우에는 0.7~5.84%가 되어 매우 정확한 결과를 나타낸다. 따라서 최초의 시공간 히스토그램 구축 시 객체의 평균 크기를 계산하여 셀 크기를 결정한다면 시공간 히스토그램은 객체의 분포와 관계없이 매우 정확한 선택도 추정이 가능한 기법이다고 할 수 있다. 그림 17에서와 같이 객체의 크기가 500이하일 경우 상대 오류율이 약간 높은 것은 객체의 크기가 균일한 경우가 분산이 큰 경우보다 정확성이 높다는 것을 의미한다.

셀 레벨 1은 셀 분할을 하지 않은 경우이며, 셀 레벨 2는 1회의 셀 분할을 통하여 x-y공간에 4개의 셀이 생성된 경우를 의미한다. 따라서 셀 레벨 1과 2에서는 셀 분할이 거의 이루어지지 않았기 때문에 셀 내에서 균일 분포가 이루어지지 못해 오류가 증가된 것이다. 그러나 셀의 크기를 객체의 평균 크기보다 작게 분할할 경우 셀 내에서 균일 분포가 보장되어 보다 정확한 선택도 추정이 가능하다.

4.4 실제 데이타 집합의 선택도 추정

실험 5에서는 GSTD 도구를 이용하여 두 개의 실제 데이타 집합을 생성하여 시공간 조인 연산의 선택도를 추정하며, 최종 상대 오류율을 계산하여 그 결과를 분석한다.

표 9 실제 데이타 집합에서 선택도 추정 실험을 위한 변수와 값

변 수	값
DENSITY	500
CENTER_AVE	6400
CENTER_VAR	1280
DISTRIBUTION	GAUSSIAN
t_q	10, 20, 30, 40, 50, 60, 70, 80, 90
i	8
SELL_SIZE(i = 1~8)	$12800/2^{i-1}$

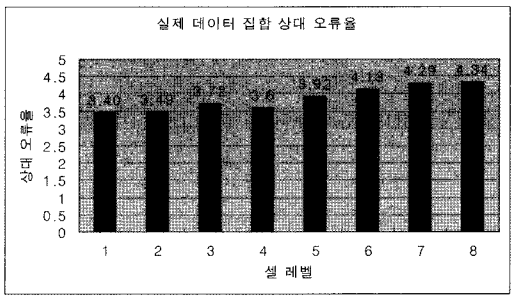


그림 19 실험 5의 상대 오류율

객체 중심점의 평균이 6400이고 분산이 1280인 가우스 분포를 이루는 데이타 집합과 임의의 중심점을 갖는 데이타 집합을 생성하였으며 객체의 밀도는 두 집합 모두 500이다. 생성된 데이타 집합에 대하여 격자 셀의 레벨을 조정하여 상대 오류율을 측정하고 각각의 셀 레벨에 대한 최종 상대 오류율은 9개의 타임스탬프에 대한 선택도 추정치의 상대 오류율의 평균값이다. 표 9는 실험 데이타의 시공간 객체가 갖는 변수 값을 나타낸다.

실제 데이타 집합에 대한 조인 연산의 최종 실험 결과는 그림 19와 같다. 그림 19와 같이 상대 오류율이 3.40~4.34%로 나왔으며, 셀 크기에 관계없이 안정적인 실험결과를 나타내고 있다. 평균점을 중심으로 편중된 분포의 모양을 갖는 가우스 분포를 이루는 데이타 집합과 임의의 중심점을 갖는 두 데이타 집합의 조인 결과는 셀 크기에 관계없이 상대 오류율이 비슷하다.

4.5 결과분석

제안한 시공간 히스토그램을 이용하여 실험 데이타와 실제 데이타 모두에 대해서 실험한 결과를 보면 셀 레벨이 3레벨 이상에서 최종 상대 오류율이 0.01~5.84%가 되어 정확한 실험 결과가 나왔다. 편중 분포를 갖는 실험 데이타에서만 셀 레벨이 1~2레벨에서 비교적 높은 오류율(18.09~25.57%)이 나왔지만 히스토그램 구축

시 격자 셀의 크기를 객체의 평균 크기로 하여 문제를 해결할 수 있다.

모든 실험에서 공통적으로 가장 정확한 히스토그램을 구축하기 위한 조건은 격자 셀의 크기를 객체의 평균 크기가 되도록 하는 것이다. 격자 셀의 크기를 객체의 평균 크기 이상으로하면 각 셀 내에서 객체가 편중 분포되어 오류율이 증가하며 객체의 평균 크기 이하로 하면 중복 계산이 되어 과추정으로 인한 오류율이 증가하게 된다.

5. 결 론

본 논문에서는 이산 데이타를 갖는 시공간 데이타베이스에서 조인 연산의 선택도 추정 기법으로 시공간 히스토그램을 제안하였다. 제안된 히스토그램을 이용하여 시간에 따라 공간 정보가 변하는 이산 데이타의 조인 연산 선택도 추정이 가능하며, 또한 시간과 공간 비용을 줄일 수 있는 히스토그램의 유지 및 관리가 가능하다. 특히 제안된 시공간 히스토그램 기법은 이산 데이타의 생명 주기가 길다는 특성을 이용하여 공간 비용을 획기적으로 줄일 수 있다. 객체가 생성 또는 소멸되는 시점을 히스토그램의 재구축 시점으로 하며, 객체의 생명주기 동안 중복되는 공간 정보를 압축, 저장하여 공간 비용을 최소화함으로써 시간 경과에 따라 히스토그램의 저장 공간 증가를 감소시켰다.

시공간 히스토그램의 정확성을 증명하기 위한 실험에서는 실험 데이타와 실제 데이타 모두를 사용하였으며 객체가 균일 또는 편중되게 분포되어 있는 경우를 모두 고려하였다. 실험결과로써 시공간 객체 분포의 균일성에 관계없이 도메인 공간을 분할하는 셀의 크기가 객체의 평균 크기보다 작으면 상대 오류율이 0.01~5.25%가 되는 높은 정확성을 보여주었다. 시공간 객체의 분포가 편중될 경우 도메인 공간을 분할하는 셀의 크기가 객체의 평균 크기보다 클 경우 상대 오류율이 다소 높게 나타나지만 셀의 크기를 객체의 평균 크기와 동일하게 하면 정확성이 유지된다. 마지막으로, 객체 분포의 균일성이 보장되는 데이타 집합에 대해서는 셀 크기와 무관하게 일정한 상대 오류율을 나타내기 때문에 보다 많은 공간 비용을 절약할 수 있다.

지리 정보 시스템, 토지 관리 시스템, 도시 계획 시스템 등과 같은 시간에 따라 변화하는 공간 정보들을 처리하는 대용량 시공간 데이타베이스 시스템에서 효율적인 질의 계획을 세우기 위해서는 선택도 추정이 중요한 역할을 한다. 이러한 데이타베이스에서 조인 연산에 대한 선택도 추정에 본 논문의 연구 결과가 활용될 수 있다.

앞으로의 연구 과제로 시공간 히스토그램에서 시간에 대한 범위 질의를 갖는 시공간 조인 연산의 선택도 추정이 가능한 히스토그램 구조에 대한 연구가 필요하다. 객체의 생명주기 동안 중복되는 객체의 압축을 통하여 공간 비용을 절약할 수 있다. 그러나 시간이 계속 경과하게 되면 히스토그램 양이 급속하게 증가하게 되므로 대용량 객체에 신속히 접근할 수 있는 자료구조에 대한 연구가 요구된다.

참 고 문 헌

[1] Hadjieleftheriou, M., Kollios, G., Tsotras, V., "Performance Evaluation of Spatio-Temporal Selectivity Estimation Techniques," In Proc. of SSDBM, 2003.

[2] Sun, J., Papadias, D., Tao, Y., Liu, B., "Querying about the Past, the Present, and the Future in Spatio-Temporal Databases," In Proc. of ICDE, 2004.

[3] Tao, Y., Sun, J. Papadias, D., "Selectivity Estimation for Predictive Spatio-Temporal Queries," In proc. of ICDE, 2003.

[4] Lee, J. and B. Shin, "Histogram-based Selectivity Estimation in Spatio-Temporal Databases," In Proc. Journal of Korea Information Processing Society, Vol.12-D, No.1, 2005.

[5] An, N., Z.Yang, and A. Sivasubramaniam, "Selectivity Estimation for Spatial Joins," In Proc of IEEE ICDE 2001, pages 368-375, Heidelberg, 2001.

[6] Belussi, A. and C. Faloutsos, "Estimating the Selectivity of Spatial Queries using Correlation Fractal Dimensions," In Proc. of VLDB '95, pages 299-310, Zurich, 1995.

[7] Faloutsos, C., B. Seeger, A. Traina, and C. Traina, "Spatial Join Selectivity using Power Laws," In Proc. of ACM SIGMOD 2000, Dallas, 2000.

[8] Theodoridis, Y. and T. Sellis, "A Model for the Prediction of R-Tree Performance," In Proc of ACM PODS '96, pages 161-171, Montreal, Canada, 1996.

[9] Theodoridis, Y., E. Stefanakis, and T. Sellis, "Cost Models for Join Queries in Spatial Databases," In Proc. of IEEE ICDE '98, pages 476-485, Orlando, USA, 1998.

[10] Mamoulis, N. and D. Papadias, "Selectivity Estimation of Complex Spatial Queries," In Proc. of SSTD 2001, pages 155-174, Redondo Beach, CA, USA, 2001.

[11] Sun, C., D. Agrawal, and A. E. Abbadi, "Selectivity Estimation for Spatial Joins with Geometric Selections," In Proc. of 8th International Conference on Extending Database Technology, EDBT 2002, pages 609-626, Prague, 2002.

[12] Belussi, A., E. Bertino, and A. Nucita, "Grid Based Methods for Estimating Spatial Join Selectivity," In Proc. of GIS'04, pages 92-100, Washington, DC, USA, 2004.

[13] Sun, J., Y. Tao, D. Papadias, and G. Kollios, "Spatio-Temporal Join Selectivity," In Proc. of

Information Systems, pages 1-21, 2005.
[14] Aref, W. and H. Samet, "A Cost Model for Query Optimization Using R-Trees," In Proc. of ACM GIS, pages 60-67, Gaithersburg, MaryLand, 1994.
[15] Kam, I. and C. Faloutsos, "On Packing R-trees," In Proc. of CIKM, pages 490-499, Washington D. C., 1993.
[16] Beigel, R. and Egemen Tanin, "The geometry of browsing," In Proc. of the Latin American Symposium on Theoretical Informatics, pages 331-340, Brazil, 1998.
[17] Acharya, S., V. Poosala, S. Ramaswamy, "Selectivity Estimation in Spatial Databases," In Proc. of ACM SIGMOD, pages 13-24, USA, 1999.
[18] Theodoridis, Y., J. R. O. Silva, M. A. Nascimento, "On the Generation of Spatio-Temporal Datasets," In Proc. of the 6th Int'l Symposium on Spatial Databases, 1999.

이 명 술

1992년 8월 전남대학교 물리교육과(이학사). 2007년 2월 충북대학교 대학원 컴퓨터공학과 석사. 2007년~현재 충북대학교 대학원 컴퓨터교육과 박사과정. 관심분야는 질의 최적화, 시공간 데이타베이스 등

이 종 연

1985년 2월 충북대학교 전자계산기공학과(공학사). 1987년 2월 충북대학교 대학원 전자계산기공학과(공학석사). 1999년 2월 충북대학교 대학원 전자계산학과(이학박사). 1989년 비트컴퓨터(주) 개발부. 1990년~1994년 현대전자산업(주) 소프트웨어연구소 주임연구원. 1994년~1996년 현대정보기술(주) CIM사업부 책임연구원. 1999년~2003년 삼척대학교 정보통신공학과 조교수. 2003년~현재 충북대학교 컴퓨터교육과 부교수로 근무중임. 2003년~2005년 한국정보처리학회 논문지 편집위원(데이타베이스분과) 역임. 2006년 한국정보처리학회 이사 역임. 2004년~현재 한국멀티미디어학회 이사와 2007년~2008년 한국산학기술학회 이사로 활동중임. 2001년 이래 IEEE Member임. 관심분야는 질의 최적화, 시공간 데이타베이스, u-learning 및 평가모델, RFID processing, GIS