시공간 점 질의를 위한 선택도 추정 기법†

신병철°, 이종연
충북대학교 컴퓨터교육과
suemirr@nate.com°, jongyun@chungbuk.ac.kr

Selectivity Estimation for Spatio-Temporal Point Queries

Byung-Chul Shin° and Jong-Yun Lee
Department of Computer Education, Chungbuk National University

요약

최근에 공간 정보들의 이력 정보를 효과적으로 다룰 수 있는 기술에 대한 연구가 활발하게 이루어지고 있다. 이러한 기술은 토지 관리 시스템이나 시간에 따라 변하는 지리 정보들을 처리하는 시스템에서 유용하게 사용되어질 수 있다. 본 논문에서는 시간의 흐름에 따라 변화하는 공간정보 질의들의 최적화를 위한 선택도 추정 기법을 제시한다. 기본 개념은 Minskew 히스토그램을 이용하여 공간 히스토그램을 구축하고, 이를 timestamp에 따라 재구축한 뒤 유지하는데 기반하고 있다. 또한, 정확한 선택도 추정률을 유지하고 히스토그램 재구축 횟수를 줄이기 위해 히스토그램 변경 내용이 최적의 임계치를 넘어 섰을 때만 시공간 데이터베이스에 현존하는 엔트리를 기반으로 히스토그램을 새로 구축하는 기법을 제시한다.

1. 서론

최근 이력 공간 데이터를 다루는 많은 애플리케이션들(예: 토지 관리 시스템, 지리 정보 시스템, 도시 계획 시스템 등)의 개발로 인하여 이력 공간 데이터를 다루는 시공간 DBMS에 대한 연구가 활발하게 진행 되고 있다. 이러한 시공간 DBMS는 시간에 따라 변화하는 공간객체들의 변화를 효과적으로 관리 할 수 있어야 한다. 이력 공간객체들에 대한 질의는 과거의 어떤 timestamp에서 특정 공간정보를 만족하는 객체들을 표현하는 것이 대표적이다. 예를 들면, "Find all objects that overlap with query area A at time t"과 같은 질의를 들 수 있다.

시공간 선택도 추정은 질의 최적화에 매우 중요하다. 시공간 선택도 추정을 하기 위해서는 실제 존재하거나 존재했던 공간 데이터를 기반으로 요약 정보를 만들고 이를 선택도 추정에 이용한다. 공간정보를 이용하여 선택도 추정을 하는 기술에는 [1], [2], [3], [4] 등 이미 많은 기술들이 나와 있다. 시공간 DBMS의 영역에는 크게 이동 객체에 대한 부분과 이력 공간정보에 대한 부분이 있다. 본 논문에서는 히스토그램을 이용하여 이력 공간 정보를 기반 한 선택도 추정을 제안하고 있다. 본 논문에서 제안하는 히스토그램의 특징은 다음과 같다. (i) Minskew 히스토그램을 timestamp 별로 구축하고 유지한다. (ii) 너무 많은 히스토그램 재구축을 방지하기 위해 히스토그램 임계치를 두어 재구축 횟수를 줄이면서 만족스러운 선택도 추정률을 유지하도록 한다. 본 논문에서 제안하는 히스토그램은 기본적으로 2차원 객체의 특정 시점에 대한 질의를 다루고 있으며 향후 연구에서 시간 범위 질의를 다룰 수 있는 히스토그램으로 확장할 것이다.

2. 관련 연구

2.1 기존의 시공간 선택도 추정

이 절에서는 이동 객체들의 이동 경로를 예측하여 선택도 추정을 하는 기존의 연구에 대하여 간략하게 기술한다.

[5]논문에서는 이동하는 객체가 고정된 질의 영역에 겹칠 수 있는지 여부에 초점을 맞추고 있다. 선택도 추정을 위하여 우선 전체 공간을 Minskew 알고리즘을 사용하여 분할하고 각 공간에 대한 선택도 추정을 한 후 이를 모두 합하여 전체 공간에 대한 선택도를 추정하는 기술을 제시하고 있다. 2차원 공간 선택도 추정은 각 차원 별로 질의와 이동 객체를 사상시켜 1차원 환경에서 선택도를 추정한 뒤 각 차원별로 구해진 선택도를 곱하여 2차원 공간에서의 선택도 추정을 한다. 따라서 그림 2.1에서 보는 것과 같이 [5]에서 제시하는 선택도 추정 기술은 2차원 공간에서는 겹치지 않는 객체가 1차원 공간으로 사상하여 겹칠 수 있는 가능성을 가지기 때문에 다차원 공간에서의 선택도 추정에 좋지 못한 성능을 보인다. 또한 히스토그램의 갱신이 너무 자주 발생하는 단점을 가지고 있다.

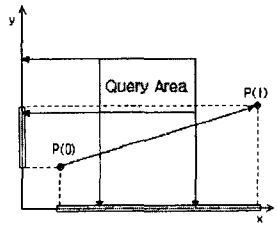
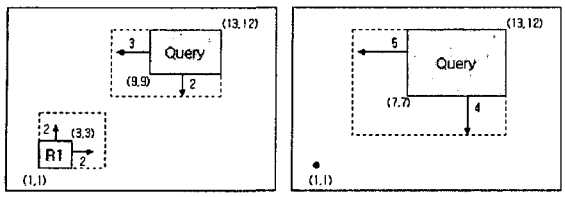


그림 2.1 초과된 선택도 추정

[6]에서는 [5]에서 제시한 이동 객체 표현을 2차원 공간으로 확장시켰다. 그리고 [5]에서의 초과 선택도 추정에 대하여 2차원 공간을 1차원 공간으로 사상시키는 것을 회피함에 따라 선택도가 높아지는 현상을 해결하였다. 어떤 객체의 현재 시간 0에서의 위치를 (x, y)라 하고 각 축에 따른 속도를 (v_x, v_y)라 할 때 [5]는 Minskew 기술을 사용하여 4차원 점 (x, y, v_x, v_y)을 표현할 수 있는 4차원 히스토그램을 생성하였다. 이 때 히스토그램의 각 버킷은 영역 MBR과 속도 MBR인 VMBR을 가지며 버킷안의 객체들은 영역 MBR과 VMBR안에서 균일하게 분포되게 히스토그램을 구축한다. 이동하는 사각형 객체와 질의에 대해서는 그림 2.2와 같이 객체들을 간소화하는 기술을 사용하여 풀어내고 있다. 히스토그램의 재구축은 전체 데이터집합에서의 갱신률이 정해진 한계 값을 초과할 경우에만 일어나므로 재구축 빈도가 [5]에 비하여 줄어들었다.

이외에도 Hough 변환과 Minskew를 이용한 [7]과 클러스터링 기술을 이용한 [8]이 있다.

† 이 논문은 2004년도 충북대학교 학술연구지원사업의 연구비 지원에 의해 연구되었음.



(a) 이동 객체와 질의 (b) 이동 객체의 간소화

그림 2.2 이동 객체의 간소화 과정

2.2 Minskew

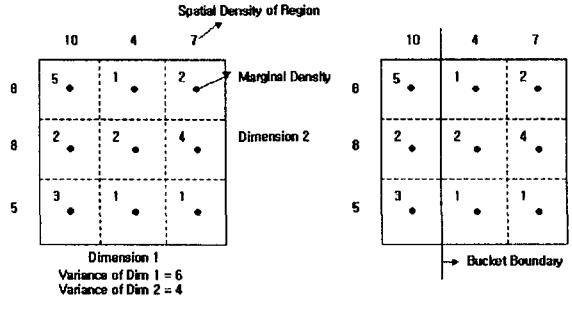
[1]에서는 공간 선택도 추정을 위한 히스토그램인 Minskew를 제안하고 있다. Minskew 히스토그램은 편중된 객체들의 버킷 분할을 통한 분배를 통하여 균일하게 만드는데 있다. 버킷 분할의 기준은 객체들의 편중도에 기반하고 있으며 분할 가능한 경우의 수중에 분할되는 두 버킷의 편중도의 가중치가 가장 낮은 분할을 선택하여 이진 공간 분할(BSP)을 한다. 각 축별로 분할 가능한 모든 경우를 검사하기에는 많은 복잡도를 가질 수 있으므로 분할 적합 축을 우선 선택한 후에 그 축을 기준으로 분할을 하는 것으로 분할 알고리즘의 효과를 높이고 있다. $B_i.num$은 i번째 버킷에 포함되는 점 객체의 수 또는 사각형 객체의 중심점이 버킷의 영역에 포함되는 수를 저장한다. C는 셀을 가리키며 $C_i.den$은 i셀에 겹치는 객체수를 저장한다. Avg(den)은 셀의 평균 den수를 말하고 |C|는 셀의 수를 가리킨다. 이 때 버킷의 편중도 $B_i.skew$는 수식 (1)과 같이 표현하며 전체 편중도의 가중치는 수식(2)로 표현한다. 최종 분할 경우의 선택은 가중치가 가장 작은 것으로 한다.

$B_i.skew$ =

$$\frac{1}{|C|} \sum_{i=1}^{number\ of\ cell\ in\ Bi} (C_i.den - Avg(den))^2 \quad (1)$$

$$Minskew = \sum_{i=1}^{n} (B_i.num \times B_i.skew) \quad (2)$$

그림 2.4의 예에서는 분할되는 축이 Dim 1이 되며 수식(1)과 (2)를 이용하여 각 분할 가능한 경우의 가중치를 구하면 첫 번째 분할이 편중도를 낮출 수 있는 방법임을 알 수 있고 그림 2.4b와 같은 결과가 된다.



(a) (b)

그림 2.4 Minskew 히스토그램의 공간 분할

3. 제안하는 기술

3.1 자료 구조

T-MinSkew 히스토그램은 2차원 공간 구조에 시간을 고려한 히스토그램이다. 히스토그램에서 시간 정보를 가지고 있는 것은 timestamp에 따른 각각의 히스토그램 밖에 없고 히스토그램내의 버킷은 묵시적으로 히스토그램의 timestamp내에 존재한다는 것을 보장한다. 그림 3.1은 간단한 T-MinSkew 히스토그램의 timestamp 0부터 12까지의 형태를 보이고 있다.

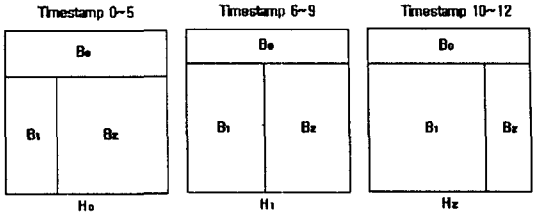


그림 3.1 T-MinSkew 히스토그램

표 3.1은 지금까지 설명한 객체들과, 질의, 버킷에 대한 기호 설명이다. 질의에는 점 질의와 영역 질의가 존재하는데 객체의 공간 영역 표현과 같으므로 표3.1에는 단순히 점 질의의 기호와 영역질의 기호만 표시하고 각 질의가 가지는 공간 영역은 표현하지 않았다.

표 3.1 자주 사용하는 기호들

기호	의미
$Qp.t$	점 질의를 하는 시간
$Qr.t$	윈도우 질의를 하는 시간
$B_i.MBR = [B_i.x_{min}, B_i.y_{min}, B_i.x_{max}, B_i.y_{max}]$	i번째 버킷의 2차원 공간 좌표
$B_i.num$	i번째 버킷이 포함하는 객체 수
$H_i.t = [H_i.t_{start}, H_i.t_{end}]$	i번째 히스토그램의 유효시간 간격
$H_i.var$	i번째 히스토그램의 객체 변화량
$timestamp_{now}$	현재시간
Sel	선택도
$OverlapArea(B_i.MBR)$	i번째 버킷과과 겹치는 질의 영역

3.2 T-Minskew 히스토그램 구축

[가정 1] 히스토그램 H_i의 시간 간격은 H_i가 생성이 된 시점부터 i+1번째 재구축이 일어난 $timestamp_{now}$ - 1 까지로 한다.

히스토그램 H_0가 실제 객체를 기반으로 timestamp 0에 구축되었을 때 유효 시간은 [0, *]을 가진다. 시간이 지나 timestamp가 1이 되었을 때 재구축을 해야 되는 지에 대한 판단을 하게 되는데 객체의 변화율과 임계치를 비교함으로써 히스토그램 재구축 여부를 결정한다. 만약 timestamp 1에서의 객체 변화율이 임계치를 넘어서지 않았을 경우는 timestamp 0에 구축된 히스토그램 H_0는 timestamp 1에도 유지가 되며 유효시간은 여전히 [0, *]로써 변화가 없게 된다. timestamp 4에서 객체 변화율이 임계치를 넘어 섰을 경우 히스토그램의 유효시간의 끝을 timestamp 3으로 기록하고 timestamp 4에는 이 시점에서 살아 있는 객체들을 기반으로 새로운 히스토그램 H_1를 생성하게 되며 유효시간을 [4, *]로 할당한다. 그림 3.2는 이러한 히스토그램 재구축 알고리즘이다.

히스토그램 재구축을 결정짓는 객체의 변화율은 히스토그램이 생성된 시점부터 실제 객체의 생성, 삭제, 갱신 횟수가 전체 객체 수에 비해 얼마나 일어났는지에 따른다. 예를 들어 전체 객체 수 N이 100이고 히스토그램 유지 기간동안의 객체 변화 횟수가 20 이라면 20%의 객체 변화율을 가지게 되며 만약 임계치가 15%라면 재구축이 일어나게 된다.

Algorithm rebuildHistogram

Step 1 $Hi.var$ += calculate variation of objects in universal space during [$timestamp_{now}$ - 1, $timestamp_{now}$]
Step 2 if ($Hi.var$ / N) over threshold
Step 3 $H_i.t_{end}$ = $timestamp_{now}$ - 1
Step 4 Create new histogram H_{i+1}
Step 5 $H_{i+1}.t_{start}$ = $timestamp_{now}$
Step 6 $H_{i+1}.t_{end}$ = *

End rebuildHistogram

그림 3.2 히스토그램 재구축 알고리즘

3.3 T-Minskew 히스토그램을 이용한 선택도 추정

선택도 추정을 위해서는 우선 질의의 timestamp와 겹치는 히스토그램을 찾은 뒤 히스토그램 내에서 질의 영역과 겹치는 버킷들을 찾는다. 질의 영역과 겹치는 각 버킷의 영역을 해당되는 버킷의 전체 영역으로 나누어 질의와 겹치는 영역의 전체 영역에 대한 비율을 구하고 이 비율에 버킷이 가지는 객체수를 곱함으로써 버킷과 겹쳐지는 버킷내의 객체수를 추정할 수 있다. 마지막으로 각 버킷에서 구해진 질의를 만족하는 객체수를 더함으로써 전체 선택도를 추정할 수 있게 된다. 수식(3)과 (4)는 위의 과정을 수식으로 나타낸 것이다.

$$Sel_j = B_j.num * OverlapArea(B_j.MBR) / area(B_j.MBR) \quad (3)$$

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

그림 3.3는 그림 3.1에서 질의의 timestamp와 겹치는 히스토그램 H_2를 가져온 것이다. $Qr.MBB$ = [5, 5, 10, 10], $B_0.MBB$ = [0, 8, 10, 10], $B_1.MBB$ = [0, 0, 7, 8], $B_2.MBB$ = [7, 0, 10, 8]이고 $B_0.num$은 3, $B_1.num$은 9, $B_2.num$은 2이다. 이 때 그림 3.4에서 각 버킷의 선택도를 추정과 전체 선택도 추정은 다음과 같이 계산된다.

Sel_1 = 1.5 = 3 * 10 / 20
Sel_2 = 0.96 = 9 * 6 / 56
Sel_3 = 0.75 = 2 * 9 / 24

$$Sel = 3.21 = 1.5 + 0.96 + 0.75 = \sum_{j=0}^{3} Sel_j$$

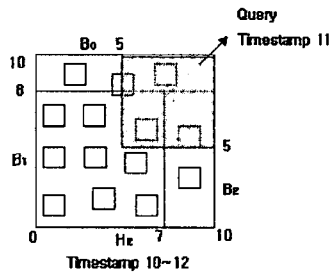


그림 3.3 T-Minskew를 이용한 선택도 추정

4. 결과 검토 및 평가

실험을 위한 상대 오류율 계산법은 수식(5)와 같다. Sel은 T-Minskew를 이용하여 질의 결과를 추정한 값이고 Sel'는 실제 질의 결과값이다.

$$Err = (Sel - Sel') / Sel' \quad (5)$$

특정 질의에 대한 편중된 결과를 벗어나기 위해 Q_n개의 질의에 대한 선택도 추정 오류율을 측정하고 평균을 구함으로써 보다 신뢰도가 높은 실험을 하도록 한다. 따라서 최종 오류율은 수식(6)과 같다.

$$Avg(Err) = (\sum_{i=1}^{Qn} Err_i) / Qn \quad (6)$$

만약 Minskew 히스토그램의 평균 오류율이 M_{err}이라면 T-Minskew의 오류율은 α만큼 증가한다. 왜냐하면 임계치 기법에 의하여 히스토그램을 유지할 때 객체의 변화에 따른 히스토그램의 재구축을 하지 않고 이전 단계에서 사용했던 히스토그램을 그대로 쓰기 때문이다. 또한 추가된 오류율 α는 높은 임계치일 수록 증가한다. T-Minskew의 오류율 TM_{err}은 수식(7)과 같다.

$$TM_{err} = M_{err} + \alpha \quad (7)$$

실험은 다음의 변수에 따라 하여 T-Minskew 히스토그램에 대한 평가를 하도록 한다.

(1) 고정된 버킷 수와 변화하는 임계에 대한 오류율
(2) 고정된 임계치와 고정된 버킷에 대한 오류율
(3) 고정된 버킷 수와 변화하는 임계치에 대한 히스토그램 재구축 횟수
(4) 고정된 임계치와 고정된 버킷에 대한 히스토그램 재구축 횟수

5. 결론

T-Minskew 히스토그램은 공간 히스토그램인 Minskew 히스토그램을 timestamp를 적용하여 이력 공간 질의에 대한 선택도 추정을 가능하게 확장한 것이다. 기본적으로 Minskew히스토그램을 사용하므로 오류율은 임계치에 따라 다소 증가할 수 있으나 전체 데이터 개수에 대한 적절한 임계치의 사용함으로써 오류율이 증가하는 것을 어느 정도 막고 정확성을 유지할 수 있다.

앞으로의 연구과제는 특정 timestamp에서의 공간 선택도 추정뿐만 아니라 범위 timestamp 선택도 추정이 가능하게 확장하고 Minskew 이외의 효과적인 timestamp 히스토그램을 구축하여 공간 선택도 추정 오류율을 낮출 예정이다.

참고문헌

[1] Acharya, S., Poosala, V., Ramaswamy, S., "Selectivity Estimation in Spatial Databases," ACM SIGMOD, USA, pages 13-24, 1999.

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

[3] Viswanath Poosala, Yannis E. Ioannidis, Peter J. Haas, and Eugene J. Shekita, "Improved Histograms for Selectivity Estimation of Range Predicates", ACM SIGMOD, NY, USA, pages 294-305, 1996.

[4] Wang, M., Vitter, J, S., Lim, L. and Pdmanabhan, S., "Wavelet-Based Cost Estimation for Spatial Queries", The 7th International Sysposium on Spatial and Temporal Databases(SSTD), CA, USA, pages 175-196, July 2001.

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

[6] Tao, Y., Sun, J., Papadias, D., "Selectivity Estimation for Predictive Spatio-Temporal Queries," ICDE, pages 417-428, 2003.

[7] Hadjieleftheriou, M., Kollios, H. and Tsotras, V J., "Performance Evaluation of Spatio-temporal Selectivity Estimation Techniques", the 15th Int. conference on Science and Statistical Database Management (SSDBM), pages 202-211, 2003.

[8] Zhan, Q. and Lin, X., "Clustering Moving Objects for Spatio-temporal Selectivity Estimation", pages 123-130, ADC 2004.