시공간데이터베이스의 다차원 선택도 추정을 위한 웨이블렛 기반 히스토그램

권정민*, 신병철, 이종연
충북대학교 컴퓨터교육과
evermin@hanmail.net○, suemirr@nate.com, jongyun@chungbuk.ac.kr

Simple Wavelet-based Histogram of Multidimensional Selectivity Estimation for Spatio-temporal Databases

Jung Min Kwon○ Byung Chul Shin and Jong Yun Lee
Department of Computer Education, Chungbuk National University

요 약

선택도 추정 기법은 상용 데이터베이스에서 질의 최적화를 위해 많이 사용하고 있다. 그 중 선택도 추정 기법에 가장 많이 사용되고 있는 기법은 히스토그램이다. 최근 시공간 데이터베이스 관련 연구에서 시간·공간 데이터베이스의 선택도 추정 기법이 활발하게 이루어지고 있다. 이 히스토그램 추정 기법이 과거에서 현재시점까지 범위 질의 수행을 성공적으로 이루어지고 있지만 대량의 데이터들을 효율적으로 관리하기에는 저장오버헤드가 너무 크다. 본 논문에서는 시공간데이터베이스에서 성공적으로 선택도 추정을 다룬 히스토그램 추정 기법을 보완하여 과거 이력데이터들의 저장을 효율적으로 할 수 있는 압축기법을 제안한다. 현재 객체에 대해서는 기존 연구에서 성공적으로 이루어진 히스토그램 기반 추정 기법을 응용하고 과거 이력데이터에 대해서는 압축기법인 웨이블렛을 응용하여 선택도추정의 오류율과 저장오버헤드의 향상이 기대된다.

1. 서 론

시공간데이터베이스에서는 공간 정보를 가지는 서열데이터와 이동객체 같은 시계열 데이터가 실생활에서 가장 많이 사용된다. 두 데이터 분야의 공통된 특징은 시간을 지원함에 따라 대량의 객체 정보를 가지는 것이다. 객체 정보량의 증가는 질의응답 시간이 증가하는 결과를 가져오기 때문에 효과적인 질의 처리를 위한 최적화 기법이 매우 중요하다. 이에 본 논문에서는 대표적인 압축기법인 웨이블렛을 통한 선택도 추정 기법을 제안한다.

1.1 연구내용 및 기여도

선택도 추정을 위한 효과적인 질의 최적화 기법에는 히스토그램 기반 기법([1,2])과 웨이블릿 기반 기법([3])이 있다. 최근 시공간 데이터베이스에서 히스토그램 기반 추정기법은 활발히 연구되었지만 대량의 공간객체 데이터 저장공간의 적절한 해결방안은 미흡하다. 최근 영상에서 성공적으로 사용되고 있는 대표적인 압축기법중 웨이블렛 기법을 데이터베이스에 적용하여 데이터를 효율적으로 압축할 수 있었다[3]. 하지만 기존 연구에서는 실시간으로 변화되고 있는 데이터에 적용할 수 있는 연구는 거의 존재하지 않는다. 또한 웨이블렛의 여러 가지 특징중의 하나는 자연스러운 다차원으로의 확장이다. 그러나 기존연구는 1차원에 대한 행과 열의 반복으로 다차원을 해결하였다. 저차원의 경우 한 차원에 대한 반복적인 수행이 가능하지만, 고차원과 시공간에서는 한 차원에 대한 단순반복으로는 신뢰성 있는 결과를 구할 수 없다[3].

따라서 본 논문에서는 데이터베이스에서 선택도 추정 기법 중 대표적으로 사용되어온 히스토그램 기반의 추정 기법과 데이터의 대표적인 압축기법인 웨이블렛을 모두 이용한 과거와 현재 시점에서의 선택도 추정 기법을 제안한다. 본 논문의 기여도는 다음과 같다. 첫째, 선택도 추정의 오류율과 저장오버헤드의 향상을 보인다. 히스토그램 기반 추정기법으로 현재 객체에 대한 효과적인 선택도 추정을 하고 웨이블렛 기반 추정기법을 함께 이용하여 과거 이력데이터에 대한 정보저장을 효율적으로 한다. 둘째, 배열개념을 이용한 차원감소를 통해 다차원을 해결한다. 다차원개념을 배열에 적용시켜 1차원으로 감소시킨 후 웨이블렛을 적용시키고, 선택도추정에 응용한다

2.관련연구

2.1웨이블렛을 이용한 다차원으로의 확장

웨이블렛을 통한 다차원에 대한 선택도 추정은 웨이블렛의 1차원 분해의 연속에 의해 이루어진다[4]. 예를 들어 2차원의 경우는 우선 데이터의 각 행(row)에 1차원 웨이블렛 변환을 한 후, 다음 같은 방법으로 각 열(column)에 대해 1차원 웨이블렛 변환을 적용한다. 각 행과 열에 대해 웨이블렛 변환을 반복함으로써 결과적으로 다차원 웨이블렛 변환을 이룰 수 있다.

2.2 2차원과 3차원에서 적용가능한 여러 가지 기법들

[6]에서는 2,3차원에 적용가능한 여러 가지 기법들을 나열하였다. Chaudhuri는 선택도추정에서 상호관계 프랙탈 차원을 사용한다. MLGF파티션기법은 동적 갱신을 제공하지만 높은 차원에서는 성능이 저하되며 실제적으로 2, 3차원에서 사용가능하지만 3차원이상에서는 적용할 수 없다.

Poosala etal이[6] 제안한 여러 가지 다차원 선택도 추정 기법중 MHIST기법이 다차원 히스토그램 기법 중 가장 좋은 결과를 나타낸다. 2차원에서 가장 낮은 에러율을 보이지만 차원의 증가는 에러율증가의 요인이 되어 3차원 이상에서는 사용할 수 없게 된다.

2.3.3 DCT(Discrete Cosine Transform)기법을 이용한 다차원으로의 확장

DCT는 이미지와 신호처리 영역에서 광범위하게 사용하는 데이터 압축 기법이다. DCT 기법의 다차원으로 일반화는 다음과 같다[LKC99].

k-차원의 DCT 계수 : $g(u_1,\ldots,u_k)=G(u(k))$, 역 DCT 계수 : $f(u_1,\ldots,u_k)=F(u(k))$

DCT변환기법의 경우 다차원에 대한 적용은 한차원에 대한 반복으로 이루어져 있다.

3.시스템 개요

실세계의 시공간 영역에 우리가 제안하는 기술을 적용하기 위하여 작업공간을 현재 살아있는 시공간 객체들과 과거에 유효했던 객체들로 분류한다. 그림 1은 간단한 시스템개요를 보여준다 .두 객체는 시간적 성질이 다르기 때문에 과거와 현재의 객체들을 각각 효율적으로 처리하기 위한 두가지 히스토그램 Current Multidimensional Histogram-based(CMH)과 Past Multidimensional Wavelet-based Histogram(PMWH)으로 나타낸다. 현재시간에 히스토그램 CMH가 재구축 되면 이전 CMH가 포함하는 객체 정보들은 유효시간의 끝 시간을 가지면서 PMWH에 저장된다. 사용자가 질의를 하면 질의의 유효시간에 따라 이력 질의와 현재 질의로 분할되어 각 질의를 담당하는 히스토그램에 넘겨진다. 질의는 공간 영역과 유효시간을 가지고 있는데 유효 시간의 시작시간 ls(live start)와 끝 시간 le(live end)가 같은 것을 시간적 점 질의(point queries with a timestamp) PQ라 하고 ls < le 인 조건을 만족하는 질의를

* 이 논문은 2005년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.

시간적 범위 질의(range queries with a time interval) RQ라 한다.

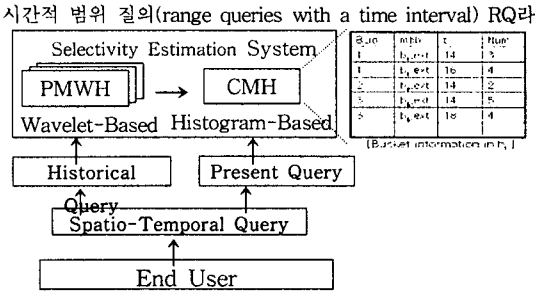
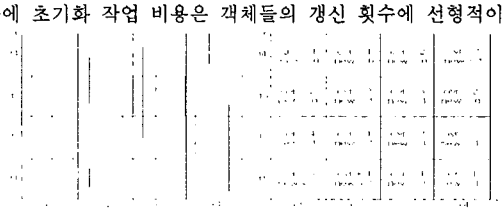


그림 1 시스템 개요

3.1 셀 초기화

전체 작업 공간은 공간영역이 w×w 크기를 가지고 시간영역이 t 크기를 가지는 셀들로 나누어진다. 버킷은 이러한 셀의 시작과 끝 인덱스를 각각 유효시간과 공간영역으로 저장한다. *bi.ovr*은 b_i 버킷과 겹치는 객체들의 수를 저장하고, *bi.new*는 b_i 버킷에서 생성된 객체들의 수를 저장한다. 이 두 변수는 시간적 범위 질의에 의한 선택도 추정의 결과에서 이전 시점부터 살아온 객체들이 결과에 중복되는 것을 막아준다

CMH는 w×w×t 개의 전체 초기화된 셀들을 기반으로 B 개의 버킷 집합을 생성한다. 예를 들어 CMH가 타임스템프 10동안 유지되었다면 이에 따른 셀의 수는 w×w×10 개가 된다. 셀 집합에는 타임스템프 10 동안 삽입, 삭제, 갱신된 모든 객체 정보들을 포함한다. 각각의 셀은 해당 셀에서 생성된 객체수를 저장하는 변수 *new*와 셀에 겹치는 객체 수 *ovr*를 저장하여 버킷을 구성할 때 근사 질의의 중복된 결과를 걸러주는 변수인 *b.ovr*과 *b.new*로 저장된다. 이러한 셀 초기화는 타임스템프마다 객체의 생성과 종료가 있는 셀에서 발생하게 된다. 그림 3은 타임스템프 now에서의 셀초기화 정보를 보여준다. 객체의 생성 및 종료와 관계없는 셀들은 셀 초기화 작업을 하지 않기 때문에 초기화 작업 비용은 객체들의 갱신 횟수에 선형적이다.



(a) (b)

그림 2 셀 초기화: (a) 객체 정보; (b) 셀 정보

예를 들어 그림 2(a)의 회색영역은 시작되는 객체 수 1개, 겹치는 객체 수가 3개이므로 2(b)처럼 $cell_{11}.new$ = 1, $cell_{11}.ovr$ = 3으로 초기화된다.

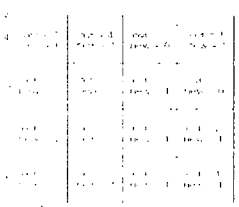


그림3 타임스템프 now에서 셀 초기화 정보

3.2 Current Multidimensional Histograms

CMH는 현존하는 객체를 위한 히스토그램기반 추정 기법이다. PQ(Point Query)를 위한 CMH는 전체 객체들을 요약하는 버킷들에 의해 구축된다. 버킷의 분할은 객체분포를 균일하게 만들기 위해 이루어진다. 현재 시간에서 초기 버킷으로부터 시작하여 B 개만큼의 버킷이 되도록 이진 분할 처리(Binary Split Processing) 과정을 거친다. 이 때 각 버킷 분할의 과정은 여러 분할 중에 원본 버킷 b_i의 객체 분산 $b_i.dis$(예 : $b_i.dis = \sum_{k=1}^{b_i.cell}(FAvg_i - c_k.ovr) / b_i.cell$)와 분할된 버킷 b_{i+1}, b_{i+2}의 평균 객체 분포 Avg($b_{i+1}.dis$, $b_{i+2}.dis$)의 차이가 가장 큰 경우를 기준으로 버킷 b_i를 분할한다. 각 버킷 분할 가치(Bucket Split Value) BSV의 계산 과정은 식(1)에 나타나 있다.

BSV = $b_i.dis$ - Avg($b_{i+1}.dis$, $b_{i+2}.dis$) 식(1)

가. CMH 구축

CMH는 셀의 초기 정보를 기반으로 BSV가 가장 큰 버킷을 이진 분할(binary splitting)한다. 버킷 분할 과정은 BST트리의 리프 노드(leaf node)의 총수가 B개가 되거나 모든 BSV 값이 음수일 때 종료한다. 그림4는 버킷의 최종분할을 보여준다.

나. CMH 버킷 갱신

CMH가 구축되고 시간이 지나 새로운 객체들의 생성과 이미 존재하는 객체들이 사라지는 경우 이에 대한 정보를 CMH내의 해당 버킷에 반영해야할 필요가 있다. 객체의 생성과 삭제가 발생한 위치를 포함하는 버킷은 기존의 객체들을 요약하는 정보를 저장하고 새로운 버킷 정보를 생성하게 된다. 기존의 버킷 정보는 새로운 버킷 정보가 저장된 위치의 주소를 가지고 있어서 선택도 추정을 할 때 BST 트리에 의해 버킷을 검색한 뒤 질의의 시간에 맞는 버킷 정보의 저장된 주소를 따라 검색하면 된다.

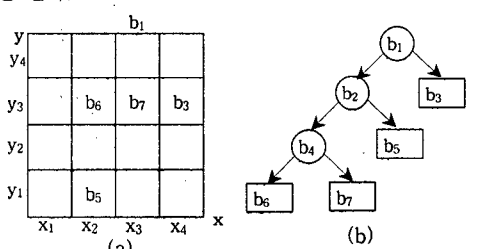


그림 4 버킷 분할의 최종 결과(b=4): (a) 분할된 버킷 정보; (b) 이진 분할 트리(BST)

3.3 Past Multidimensional Wavelet-based Histogram

시간이 지남에 따라 CMH에 있는 객체 분포들은 히스토그램 생성시 균일한 성질을 잃어버리게 되고 객체 분포의 분산이 주어진 임계치를 넘어설 경우 객체 분포의 균일 성질을 유지하기 위해 히스토그램을 재구축한다. CMH 재구축에 의해 분리된 객체 집합은 CMH가 유지되는 동안에 변화된 객체 정보를 기록한 3차원 셀 배열에 저장되어 있다. PMWH 구축은 이 3차원 셀 배열에 저장되어 있는 정보를 기반으로 이루어지며 셀의 시간정보를 포함한다. PMWH는 CMH와는 다르게 히스토그램 기반이 아닌 웨이블렛 압축기법을 기반으로 구축된다. PMWH는 객체정보들이 웨이블렛 변환을 통해 웨이블렛 계수로 산출되고 임계치 과정을 거친 후 웨이블렛 계수 형태로 저장된다.

가. PMWH구축

CMH에서는 버킷의 갱신으로 각 셀에 대한 객체정보를 다루었지만 PMWH는 버킷이 아닌 셀의 개념으로 구축된다. 각 셀은 셀 자신이 포함한 객체에 대한 객체수를 가지고 있다. PMWH의 구축은 시간정보를 저장한 시간차원과 2차원공간이다. 먼저 배열을 이용하여 2차원 배열인 2D공간을 1차원으로 감소시킨다. 1차원으로 감소된 각 셀의 객체에 대한 정보들은 웨이블렛 기법을 통해 웨이블렛 계수로 변환되어 구축된 시점 now를 기준으로 아래의 그림 6과 같이 시간정보와 함께 PMWH에 저장된다.

4. 선택도 추정

시공간 질의에 대한 선택도 추정은 구축된 PMWH와 CMH를 이용한다. 선택도 추정을 위해 사용되는 추정기법은 두 질의(PQ, RQ)의 따라 해당되는 PMWH와 CMH를 선택하게 된다.

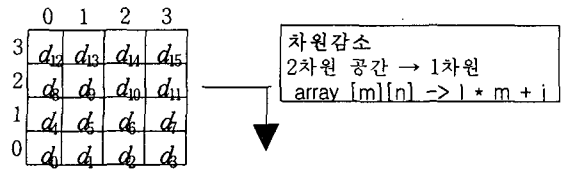


그림 5. 2차원공간의 차원감소

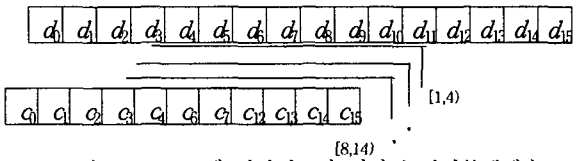


그림 6. PMWH에 시간정보와 저장된 웨이블렛계수

4.1 CMH에서의 선택도 추정

가. 시간적 점 질의

점 질의 PQ의 공간 정보와 겹치는 CMH내의 버킷들을 모두 검색한다. 질의 공간 영역과 겹치는 각 버킷의 공간 영역을 해당되는 버킷의 전체 공간 영역으로 나누어 질의와 겹치는 공간 영역의 전체 공간 영역에 대한 비율을 구하고 이 비율에 버킷이 질의 시간에 가지는 객체수를 곱함으로써 질의와 겹쳐지는 버킷내의 객체수를 추정할 수 있다. 끝으로 각각 구한 추정된 선택도를 모두 더함으로써 CMH내의 PQ에 대한 전체 선택도 추정을 구할 수 있다. 식 (2)와 (3)에 이러한 과정을 나타낸다.

$Sel_j = b_j.ovr * OverlapArea(b_j)/area(b_j)$ 식(2)

$$Sel = \sum_{j=0}^{k} Sel_j$$ 식(3)

나. 시간적 범위 질의

시간적 범위 질의 RQ의 선택도 추정은 PQ의 선택도 추정을 확장한다. RQ를 위한 선택도 추정은 PQ와는 달리 질의가 범위 시간을 가지고 있기 때문에 보통 히스토그램에서는 시간적으로 하나의 객체가 여러 버킷에 걸쳐 있을 수 있어 선택도 추정의 결과에 객체가 중복될 수 있다. 이를 위한 해결책으로 히스토그램 내의 각 버킷에 할당된 new와 ovr 변수를 이용하여 중복된 객체 적용을 피할 수 있다.

RQ에 대한 선택도 추정은 식 (2)를 적용하여 사용할 수 있고, 각 버킷에 대한 전체 선택도는 식(3)을 이용하여 추정할 수 있다.

4.2 PMWH에서의 선택도 추정

시간정보와 함께 저장된 과거 이력데이터인 PMWH에 대한 선택도 추정은 그림 7과 같이 웨이블렛계수 형태로 저장되어 있기 때문에 초기값으로의 복원과정이 필요하다.

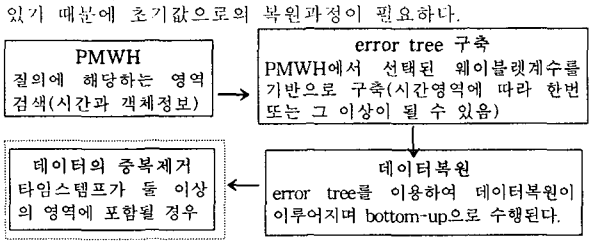


그림 7 PMWH에서의 데이터 복원과정

PMWH에서도 두가지 PQ와 RQ의 두 가지 경우가 발생한다. PQ의 경우는 웨이블렛의 데이터 복원과정으로 간단하게 추정할 수 있다[3].

시간 범위 질의 RQ는 여러 가지의 경우가 발생할 수 있다. 첫째, 타임스템프가 하나의 웨이블렛 계수 저장영역에 해당될 경우이다. 그림 6에서 영역질의가 타임스템프 8~10 사이인 경우가 된다. 가장 먼저 시간 정보를 검색하여 웨이블렛 계수가 저장되어 있는 영역을 검색한다. 타임스템프가 [8,10)인 웨이블렛 계수가 선택된다. 선택된 웨이블렛 계수를 기반으로 error tree를 구축되고 웨이블렛 복원방법으로 데이터복원이 이루어진다[3]. 복원된 추정값이 선택도가 된다. 둘째, 타임스템프가 두 개이상의 저장된 웨이블렛의 영역을 포함할 경우이다. 예를 들어, 타임스템프가 7~10사이라는 범위가 정해졌을 때는 두 개의 웨이블렛 계수 영역이 필요하게 되고, error tree 구축이 두 번 이루어져야 한다. 복원된 데이터값의 합이 전체선택도 추정값이 되는데 여기에서의 문제점은 객체의 중복이다. CMH에서 PMWH로 저장될 당시 CMH 재구축 시점에서의 객체정보를 그대로 반영하여 저장한 것이므로 PMWH에 저장된 데이터들은 서로 중복될 가능성이 높기 때문에 이 경우 중복제거의 과정이 추가로 필요하게 된다. 셋째, 시간 범위가 CMH와 PMWH에 모두 적용될 경우로 CMH선택도 추정의 식(2)와 PMWH에서 첫 번째경우와 같은 과정을 통해 얻어진 복원값을 더한것이 전체 선택도 추정값이 된다.

5. 결과 분석

실험을 위한 상대 오류율 계산법은 식 (4)와 같으며 *Sel*은 히스토그램과 웨이블렛 기법을 이용한 질의 추정 값이고 *Sel'*는 실제 질의 결과 값이다.

$Err = (Sel - Sel') / Sel'$, 단 $Sel'>0$ 식(4)

특정 질의에 대한 편중된 결과를 해결하기 위해 Q_n개의 다수 질의에 대한 선택도 추정 오류율을 측정하고 그것의 평균을 구함으로써 실험에 보다 높은 신뢰도를 가지도록 한다. 최종 오류율을 구하는 식은 (5)와 같다.

$Avg(Err) = (\sum_{i=1\sim N} Err_i) / N$ 식(5)

실험은 시간적 점 질의와 범위 질의에 대해 이루어진다. 히스토그램 기반으로 추정한 선택도 추정값과 웨이블렛기반 추정 기법에 대해 [5]의 연구와 비교 평가한다. 이 비교실험은 히스토그램기반으로 CMH와 PMWH 모두를 구축 하였을 경우와 과거 이력데이터에 대해 PMWH를 적용하였을 경우 선택도 추정의 오류율과 저장공간에 대한 오버헤드에 대한 향상이 기대된다.

6. 결론

본 논문에서는 시공간 데이터베이스에서 서열 데이터를 위한 선택도 추정 기법으로 한 히스토그램 CMH와 웨이블렛 기반 추정 기법을 적용한 PMWH를 제안하였다. 제안된 두 개의 추정기법으로 과거와 현재 까지의 공간 객체들에 대한 효과적인 선택도 추정을 할 수 있다. 본 논문에서 제안한 기법은 기존 연구와 비교하여 히스토그램 기반으로 저장된 PMH[5]에 비해 저장오버헤드를 향상시킬 수 있고, 기존 차원의 반복으로 인한 다차원 해결이 아닌 배열을 이용한 차원감소를 통해 다차원을 해결함으로써 효과적인 선택도 추정을 할 수 있다.

앞으로의 연구 과제는 PMWH에 사용된 웨이블렛 기법의 임계치을 향상시켜 저장되지 않은 웨이블렛 계수가 초기 데이터 복원값에 대한 미치는 오류율을 감소시킴으로써 선택도 추정 오류율을 줄이며, 현재 히스토그램 기반으로 이루어진 선택도 추정을 효과적으로 웨이블렛에 적용시킬 수 있는 알고리즘을 개발하는 것이다.

참고문헌

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

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

[3] Yossi Matias, Jeffrey Scott Vitter, and Min Wang, "Wavelet-Based Histogram for Selectivity Estimation," In Proceedings of ACM SIGMOD international conferences on Management of data, pages 448-459, 1998.

[4] Y.Matias, J.S. Vitter, and M. Wang. "Dynamic Maintenance of Wavelet-Based Histogram ". In *Proc. of the 26th INtl. Conf. on Very Large Data Bases*, September 2000.

[5] Shin, B. and Lee, J., "Selectivity Estimation for Multidimensional Sequence Data in Spatio-Temporal Databases", In Korean DataBase Conference, pages 160-166, May, 2005

[6] J.Lee, D.Kim, and C.Chung. "Multi-dimensional selectivity estimation using compressed histogram information. In *Proceedings of the 1999 ACM SIGMOD internation Conference on Management of Data*, pages 205-214, Phildelphia, June 1999.