다차원 히스토그램을 이용한 공간 위상 술어의 선택도 추정 기법

김 홍 연† · 배 해 영††

요 약

히스토그램을 이용한 질의 술어의 선택도 추정은 상용 데이터베이스 시스템의 비용 기반 최적화기에서 가장 널리 사용되는 방법이다. 공간 데이터베이스 관리 시스템의 경우 객체간의 위상 관계를 이용한 술어가 주어지며, 질의 최적화를 위해서는 공간 위상 술어의 선택도 추정이 필수적이다. 이를 위해 본 논문에서는 기존의 다차원 히스토그램 기법에 차원 변환 기법을 적용한 공간 위상 술어 추정 기법을 제안한다. 제안된 기법은 차원 변환 공간 상의 점으로 대응된 공간 객체로부터 두 가지 분할 전략을 이용하여 공간 히스토그램을 생성한 후 변환 공간이 가지는 위상 관계를 이용하여 공간 위상 술어의 선택도를 추정한다. 제안된 기법은 공간 질의 최적화기에서 비교적 작은 메모리와 부가적인 입출력 없이 공간 위상 술어의 선택도를 추정할 수 있다.

A Selectivity Estimation Scheme for Spatial Topological Predicate Using Multi-Dimensional Histogram

Hong-Yeon Kim† · Hae-Young Bae††

ABSTRACT

Many commercial database systems maintain histograms to summarize the contents of relations, permit efficient estimation of query result sizes, and access plan costs. In spatial database systems, most query predicates consist of topological relationship between spatial objects, and it is very important to estimate the selectivity of those predicates for spatial query optimizer. In this paper, we propose a selectivity estimation scheme for spatial topological predicates based on the multi-dimensional histogram and the transformation scheme. Proposed scheme applies two partition strategies on transformed object space to generate spatial histogram, and estimates the selectivity of topological predicates based on the topological characteristic of transformed space. Proposed scheme provides a way for estimating the selectivity without too much memory space usage and additional I/Os in spatial query optimizer.

1. 서 론

RDBMS는 SQL 등의 선언적인 방법으로 주어진 질의를 가장 작은 비용으로 처리할 수 있는 방법을 자동으로 산출해 낸다는 특징을 가지고 있다. 질의 처리 비용을 정확하게 산출하는 것은 매우 복잡하지만[1,2,3] 질의를 처리하기 위해 수행되어야 하는 I/O의 횟수에 비례한다고 알려져 있다[4,5]. 질의 처리를 위해 필요한 I/O는 질의 처리 과정에서 참조되어야 하는 튜플들의 개수 및 질의 트리 상의 중간 결과 튜플들의 개수와 밀접한 관련을 가지고 있다. 질의 처리 비용을 최소화

† 준 회 원 : 인하대학교 대학원 전자계산공학과
†† 종신회원 : 인하대학교 전자계산공학과 교수
논문접수 : 1998년 11월 12일, 심사완료 : 1999년 2월 19일

하는 것은 질의 트리의 각 단계에서 참조되어야 하는 튜플들의 개수 및 중간 결과 튜플의 크기를 최소화 하는 것이라고 할 수 있으며, 이를 위해서는 질의 트리의 각 단계에 명시된 질의 술어를 만족하는 튜플 개수의 정확한 추정이 필수적이다[4].

질의 결과를 추정하기 위하여 사용되는 대표적인 방법은 데이터베이스에 저장되어 있는 릴레이션에 다양한 종류의 통계 수치들을 유지하는 것이다. 이러한 통계 수치들은 릴레이션의 각 속성들에 저장되어 있는 데이터 값의 실제 분포를 추정하기 위한 수치들이다. 질의 처리 결과 크기 추정[5]을 위해 제안된 방법으로서 대표적인 것으로 히스토그램[6], 샘플링[7,8], 그리고 파라메트릭 기법[9,10] 등이 있다. 이중 히스토그램 기법은 릴레이션의 속성의 현재 데이터 값의 빈도 및 분포를 버켓 이라고 불리는 곳에 유지하고 이를 이용하여 주어진 질의 술어의 조건을 만족하는 데이터의 개수를 추정하기 위해 사용한다. 다른 기법에 비해 히스토그램 기법이 가지는 장점은 시스템 카탈로그 상에 비교적 작은 크기의 히스토그램을 유지하는 것만으로 질의 실행 시에 별다른 부가 노력을 필요로 하지 않고 질의 처리 결과 크기를 추정할 수 있다는 점을 들 수 있다[11]. 따라서 히스토그램 기법은 상용 DBMS(DB2, Informix, Ingres, MS-SQL, Sybase)에서 가장 많이 채용하고 있는 기법이다.

최근 들어 이러한 기법을 지리정보 시스템, CAD/VLSI 데이터베이스 시스템 등에 적용하기 위한 연구가 진행중이다[12,13,14]. 공간 질의 최적화 역시 비공간 질의 최적화와 같이 각 질의 처리 단계에서 참조되는 튜플들의 크기 및 중간 결과물의 크기를 정확히 산출하여야 하며 상대적으로 비공간 질의에 비해 공간 질의 처리 비용이 매우 크므로 잘못된 질의 처리 비용 추정은 시스템 성능에 매우 치명적인 영향을 끼친다.

공간 술어 결과의 크기를 추정하기 위해서는 공간 데이터가 두 개 이상의 좌표 값으로 구성된 복합 객체의 특성을 가지므로 기존의 히스토그램 기법, 샘플링 기법, 파라메트릭 기법 등을 단일 속성에 대한 통계 기법에서 두개 속성 이상의 통계 기법으로 확장하여야 한다. 이러한 확장 기법으로 대표적인 것으로 다차원 질의 결과 추정을 위한 히스토그램을 이용한 선택률 추정 기법[12,13]과 다차원 동적 화일을 이용한 선택률 추정 기법[14] 등이 있다. 전자의 경우 n개로 이루어진 복합 객체를 n차원 상의 점에 대응시키고 이 공간을 여러 개의 버켓으로 분할(partition)하는 추정 기법이다. 그러나 이 기법을 그대로 공간 객체의 최소경계다각형(MBR, minimum bounded rectangle)에 적용 시켰을 경우 MBR이 4차원상의 단일 점으로 매핑 되어 MBR이 가지는 영역 개념 및 MBR 간의 위상 관계를 잃어 버리는 문제점을 가지고 있다. 또한 후자의 경우 공간 색인 기법의 하나인 차원 변환 기법과 다차원 동적 화일에 기반하여 공간 데이터의 분포 통계 수치를 유지하기 때문에 공간 데이터의 갱신 시 자동적으로 통계수치가 갱신된다는 장점을 가지고 있는 반면 히스토그램 기법과는 달리 다양한 분할 전략을 통한 분포상 편이(skew) 현상에 대처하지 못한다는 단점을 가지고 있다. 그러나 이 기법에서 사용하고 있는 차원 변환 기법을 히스토그램 구축에 적용한다면 공간 색인 기법에 의존적이지 않는 히스토그램의 구축이 가능하다.

본 논문에서는 이러한 고찰을 기반으로 히스토그램에 기반한 공간 자료의 위상 관계 술어의 선택률 추정 기법을 다룬다. 기존의 다차원 히스토그램이 MBR의 특성을 손실하는 문제를 해결하기 위하여 공간 색인 기법에서 사용되는 변환 기법을 적용한 추정 기법을 제안하며, 이 기법이 공간 질의 술어의 선택률 추정에 효과적으로 적용 될 수 있음을 보인다. 또한 히스토그램에 기존의 다양한 분할 전략을 적용하여 실제 공간 데이터의 선택률 및 선택률 오차의 변화를 실험한다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 다차원 히스토그램이 MBR로 표현되는 공간 객체간의 관계를 잃어 버림을 보이고 이 특성이 기존의 차원 변환 기법에 의해 극복될 수 있음을 보인다. 또한 차원 변환 공간에 기존의 분할 전략을 적용하여 히스토그램으로 쉽게 변경할 수 있음을 보인다. 3장에서는 이를 기반으로 공간 위상 술어의 선택도 추정을 위한 특성 및 기법을 다루고 4장에서는 제안된 기법에 대한 성능 평가를 다룬다. 끝으로 5장에서 결론을 맺는다.

2. 공간 데이터 분포 공간의 특성

기존 히스토그램 기법은 선택도 추정 시 오류를 최소화하기 위하여 데이터 값의 분포 공간을 equal width, equal depth[4,6,12], variable width[18], serial [19, 20,21] 등의 다양한 분할(partition) 정책을 사용하여 분할 한 후 각 분할 영역에 통계 수치들을 저장한다. 이들 히스토그램 기법에서 대상이 되는 데이터는 문자

열, 숫자 등 단순한 자료형을 기본으로 하고 있으므로 도메인 상의 한 점으로 대응될 수 있다. 따라서 이러한 데이터의 분포는 다양한 분할 규칙에 의하여 단일 분할 영역으로 나뉘어 질 수 있다.

공간 데이터의 선택도 추정을 위해 히스토그램을 구축하기 위해서는 공간 데이터의 분포를 기존의 여러 분할 기법을 사용하여 분할하고 통계 값을 유지하여야 한다. 그러나 기존 속성 데이터들이 도메인상의 한 점에 대응되는 것과 달리 공간 데이터는 도메인상의 영역으로 대응된다. 따라서 속성 데이터를 위한 분할 기법과는 달리 공간 데이터의 분포 공간을 그대로 분할할 경우 하나의 공간 데이터가 여러 개의 분할에 걸쳐 존재하게 되는 문제가 발생한다.

이러한 문제는 공간 데이터가 영역을 가지고 있기 때문에 발생하는 것으로서 R 트리[15], Grid 트리[16] 등 공간 색인 기법에서도 공통적으로 발생한다. 이 문제는 2차원상의 MBR(x1,y1,x2,y2)을 4차원 공간상의 한 점에 매핑하여 해결할 수 있다. [12]는 다차원 공간을 equal depth 분할 하기 위한 알고리즘과 히스토그램 저장구조 그리고 half scheme과 uniform scheme의 두 가지 추정 기법을 제안하였다.

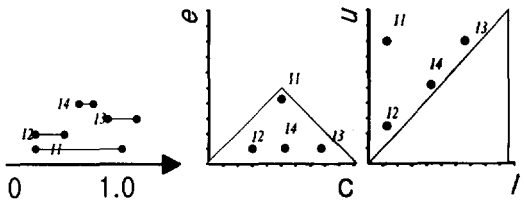
제안된 추정 기법은 두개의 n차원상의 점을 사용하여 지정된 질의 영역의 내부에 포함되는 데이터의 개수를 추정한다. 이 기법을 지정된 영역과 겹치는 공간 객체를 선택하기 위한 다음과 같은 공간 술어에 적용시킬 경우 두 가지 문제가 발생한다.

Find all building that overlap rect (x1, y1, x2, and y2)

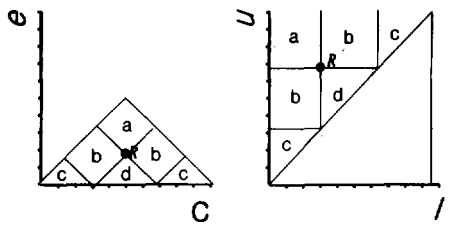
첫째, 검색의 대상이 되는 공간 객체와 질의 영역을 지정하기 위한 MBR의 차원 및 점 개수가 동일하기 때문에 두개의 동일 차원상의 점을 사용하는 기존의 추정 기법을 사용할 수 없다. 둘째, 공간 객체와 MBR 간의 겹침 위상 관계에 의한 선택도 추정이 불가능 하다. 공간 질의에서는 겹침 이외에도 다양한 위상 관계에 의한 술어가 필수적이므로 공간 술어를 위한 추정 기법의 확장이 필요하다.

이 문제는 기존 공간 색인 기법 중 차원 변환 기법(T-schemes)[17]의 특성을 사용하여 해결할 수 있다. 이 기법에는 크게 중앙 점 변환 기법(center T-scheme)과 구석 점 변환 기법(cornet T-scheme)이 있으며, n차원상의 공간 데이터의 MBR을 2n 차원상의 한 점으로 대응시키는 기법이다. (그림 1)은 1차원상의 선분들이 두 가지 차원 변환 기법에 의해 2차원상의 점으로 변환됨을 보인다. 이 기법은 (그림 2)과 같이 MBR R에 대응되는 지점을 중심으로 각 공간 위상 관계가 명확히 구분된다는 특징을 가지고 있다. 이 기법을 사용하여 구분 가능한 위상 관계는 공간 질의 시 필요한 동일(지점R), 포함(영역d), 포함됨(영역a), 겹침(영역b), 인접(b-c경계), 떨어짐(영역c)의 6가지 모두가 표현된다.

변환 공간의 또 다른 특징은 현재 데이터베이스에 저장되어 있는 모든 공간 객체의 크기 및 위치에 대한 개괄적인 요약 정보를 제공한다는 것을 들 수 있다. 중앙 점 변환 기법의 경우 c축, 구석 점 변환 기법의 경우 대각선상에 근접한 점이 많을 수록 작은 크기의 공간 객체가 많음을 의미하고, 반대의 경우 큰 공간 객체가 많음을 의미한다. 또한 c축 및 대각선의 좌, 우측에 위치하는 공간 객체는 원 공간에서도 좌, 우측에 존재함을 의미한다. 변환공간의 이러한 특징은 공간 데이터베이스의 통계적 상태를 나타내기 위한 매우 적절한 특성이다.

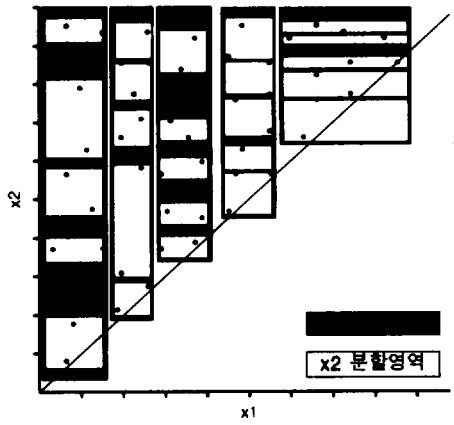


(그림 1) 차원 변환 기법
(Fig. 1) Transformation Schemes



(그림 2) 변환기법과 위상 영역
(Fig. 2) T-Scheme and topology

이 같은 변환공간은 (그림 3)과 같이 다차원 히스토그램 기법에서 제안된 equi-depth 분할 알고리즘[12]을 적용하여 히스토그램으로 변환될 수 있다.

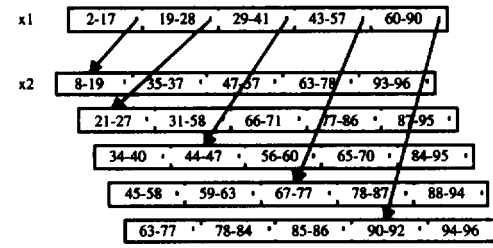


(그림 3) Equi-Depth 분할의 예
(Fig. 3) Example Equi-Depth Partition

Algorithm GenericP는 [12]에서 제안된 변환 공간을 분할하기 위한 개략적인 알고리즘이다.

```
Algorithm Node GenericP(Col, Row, Rows)
Var  i, i0
     node = MakeNode()
     Sort(Col, Row, Row + Rows - 1)
     i = Row
     While  i < Row + Rows
          i0 = i
          i = FindBoundary()
          lb = V(i0, Col),  ub = V(i-1, Col)
          If  Col < 4  Then
               hild = GenericP(Col + 1, i0, i - i0)
               Insert(node, lb, ub, child)
          Else
               f = i - i0
               nsert(node, ib, ub, f)
          End
     End
     GenericP = node
End
```

알고리즘 GenericP는 다차원 공간의 분할을 위한 일반화된 알고리즘으로서 버켓의 경계를 구하기 위한 FindBoundary 알고리즘을 변경하여 equi-width, equi-depth, v-optimal 등의 다양한 분할 정책을 적용 할 수 있다. 알고리즘 GenericP를 이용하여 생성한 히스토그램 트리는 (그림 4)와 같다.



(그림 4) GenericP가 생성한 히스토그램 트리
(Fig. 4) A histogram tree generated by GenericP

생성된 트리는 차원 개수와 동일한 단계를 가지고 있다. 루트 노드는 하나의 노드를 가지며 검색의 시작점이 된다. 각 단계는 변환 공간의 각 좌표축을 의미하며, 각 단계는 하나 이상의 노드들로 구성된다. 각 노드는 여러 개의 슬롯으로 구성되며, 각각의 슬롯은 해당 좌표축에서의 분할 범위와 다음 단계로의 링크를 위해 최소값, 최대값, 링크로 구성된다. (그림 4)의 경우 2차원 변환 공간을 히스토그램 트리로 표현한 것이다. MBR의 경우 4차원으로 변환 되므로 네 단계로 이루어진 히스토그램 트리가 생성된다.

3장에서는 이러한 특성을 바탕으로 변환 공간에 기반한 히스토그램을 사용한 공간 술어의 선택도 추정 기법을 제안한다.

3. 공간 술어 선택도 추정 기법

3.1 변환 공간상에서의 선택도 추정

공간 술어는 질의 영역으로 주어진 MBR과 특정 위상 관계에 있는 모든 튜플을 검색하기 위한 조건이다. 공간 술어의 선택도를 추정하기 위해서는 변환 공간상에 구축된 히스토그램의 모든 분할 중에서 질의 영역과 주어진 위상 관계를 만족하는 분할을 검색해야 한다.

공간 술어에 명시된 위상 관계를 만족하는 분할은 (그림 2)의 각 위상 관계별 대상 영역과 겹치는 분할이다. 이에 따라 각 분할은 주어진 질의 영역 및 위상 관계에 대응되는 영역과 완전히 겹칠 수도 있고 일부분만 겹칠 수도 있다.

정의 1 : f-버켓은 주어진 위상 관계 영역과 완전히 겹치는 버켓이다.

정의 2 : p-버켓은 주어진 위상 관계 영역과 일부분이 겹치는 버켓이다.

주어진 질의 영역을 만족하는 버켓들이 모두 f-버켓이라면 추정 오류는 존재하지 않는다. 만약 주어진 질의 영역을 만족하는 버켓들에 p-버켓이 포함되어 있으면 버켓 내의 튜플들에 대한 균일 분포 가정(uniform distribution assumption)에 의해 추정 오류가 발생한다.[12] f-버켓의 튜플 수를 $F_i(0 \le i < f-1)$, p-버켓의 튜플 개수를 $P_j(0 \le j < p-1)$라고 한다면 주어진 질의 영역 및 위상 관계를 만족하는 실제 튜플의 개수 N_q의 범위는 <수식 1>과 같다.

$$\sum_{i=0}^{f} F_i \le N_q \le \sum_{i=0}^{f} F_i + \sum_{j=0}^{p} P_j \qquad \text{<수식 1>}$$

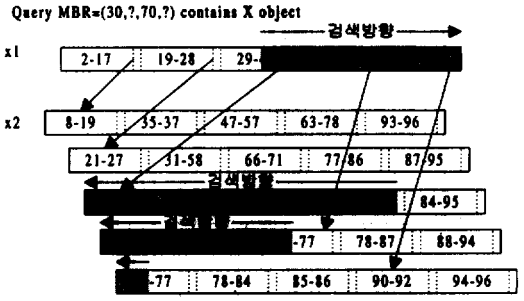
<수식 1>은 uniform scheme과 half scheme[12]에서 제시한 오류의 범위를 f-버켓과 p-버켓을 이용하여 일반화 한 것이다. 추정 시 발생 하는 오류는 p-버켓에 의한 것이며 주어진 위상 관계를 만족하는 튜플의 개수를 정확히 추정하기 위해서는 히스토그램의 각 버켓에서 f-버켓과 p-버켓을 정확히 판별해 내야 한다.

3.2 위상 관계에 따른 히스토그램 검색

주어진 위상 관계에 대해서 f-버켓과 p-버켓을 판정하기 위해서는 (그림 2)의 위상 관계별 대응 영역으로부터 유도된 <표 1>의 관계를 사용하여 히스토그램 트리를 검색한다. <표 1>에서 x_1, x_2, y_1, y_2는 히스토그램 트리의 해당 단계의 값들이며, mx_1, mx_2, my_1, my_2는 질의 영역으로 주어진 MBR이다. 예를 들어, 질의 영역 (30, 40, 70, 80)에 포함되는 버켓을 찾기 위해서는 (그림 7)과 같이 히스토그램 트리의 최상위 노드인 x_1 단계에서 30보다 큰 값을 가지는 슬롯들을 검색한다. 선정되는 슬롯들은 (그림 5)에서 회색 영역을 포함하는 슬롯이며, 슬롯에 저장된 링크로부터 x_2 단계 노드를 구한다. 이 과정에 의해 3개의 x_2 단계 노드가 선별되며, 다시 70보다 작은 값을 포함하고 있는 슬롯이 가리키는 y_1 단계 노드를 선별한다. 선별된 y_1 단계 노드에 대해서도 <표 1>에 근거하여 동일한 과정을 반복하면 y_2 단계의 노드들이 선별된다. y_2 단계의 노드는 히스토그램 트리 상의 잎 노드이며 이 단계에서 80보다 작은 값을 가지는 슬롯들을 구한다. 이 과정을 거쳐 검색된 y_2 단계의 슬롯들은 변환 공간 상에서 주어진 위상 관계를 만족하는 버켓들이며, 다른 노드로의 링크 대신에 해당 버켓의 튜플 수를 저장한다.

<표 1> MBR과 버켓간의 위상관계
<Table 1> Topological relationship between MBR and Bucket

위상관계	x_1	x_2	y_1	x_2
동일	mx_1=	mx_2=	my_1=	my_2=
포함	mx_1<	mx_2>	my_1<	my_2>
포함됨	mx_1>	mx_2<	my_1>	my_2<
인접	mx_2=			
		mx_1=		
			my_2=	
				my_1=
겹침	mx_1< and mx_2>		not my_2<	not my_1>
		mx_1< and mx_2>	not my_2<	not my_1>
	not mx_2<	not mx_1>	my_1< and my_2>	
	not mx_2<	not mx_1>		my_1< and my_2>
disjoint	mx_2<			
		mx_1>		
			my_2<	
				my_1>



(그림 5) 포함 관계의 추정
(Fig. 5) Estimation of cover relationship

검색된 버켓들은 f-버켓 및 p-버켓을 모두 포함한다.

정의 3 : f-슬롯은 검색 영역을 완전히 포함하는 슬롯이다.

정의 4 : p-슬롯은 검색 영역을 일부분 포함하는 슬롯이다.

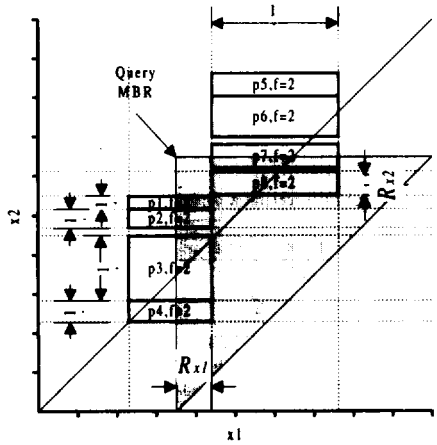
검색된 버켓이 f-버켓이기 위해서는 부모 슬롯들이 모두 f-슬롯이어야 하며, 그렇지 않은 경우는 p-버켓이다.

3.3 공간 술어 선택도 추정 기법

공간 술어의 선택도 추정 시 발생하는 오류는 위상 관계 영역이 버켓을 완전히 포함하지 못하는 p-버켓으로부터 발생한다. p-버켓에 의한 오류는 기존의 속성 데이터를 위한 다수의 히스토그램 기법에서도 존재하며 이것은 버켓 내부의 튜플들은 균일 분포를 따른다는 가정에 기인한 것이다.

균일 분포 가정(uniform distribution assumption)에 따른 추정 기법에서는 질의 영역과 버켓이 겹치는 비율에 따라 해당 버켓의 튜플 개수를 조절하여 반영한다.

해당 버켓과 위상 관계 영역이 겹치는 비율은 (그림 6)에서 각 단계의 p-슬롯이 검색영역과 겹치는 비율로부터 얻어낼 수 있다.



(그림 6) p-버켓 및 추정 인수
(Fig. 6) p-bucket and estimation parameters

각 버켓 내부의 분포는 균일 분포를 따르므로 포함 위상 관계의 경우 다음과 같은 수식에 의해 x_1, x_2, y_1, y_2의 네 단계의 p-슬롯들의 비율을 계산할 수 있다.

$$R_{x1} = \frac{ub - mx_1}{ub - lb},\ R_{x2} = \frac{mx_x - lb}{ub - lb}$$

$$R_{y1} = \frac{ub - my_1}{ub - lb},\ R_{y2} = \frac{my_2 - lb}{ub - lb}$$

R_{x1}, R_{y1}는 위상 관계 영역 경계선상의 우측이 겹치는 부분이고, R_{x2}, R_{y2}는 경계선상의 좌측이 겹치는 부분이다. 포함 관계 이외의 비율은 모두 <표 1>로부터 유도해 낼 수 있다. 네 가지 비율 모두 0에서 1사이의 값을 가지며 1일 경우 f-슬롯을 의미하며 1미만일 경우 p-슬롯임을 의미한다. (그림 7)은 mx_1, mx_2 만을 이용하여 이차원 상에서 포함 관계 질의 시 R_{x1}과 R_{x2}을 표시한 것이다.

이와 같은 비율을 기반으로 f가 현재 버켓의 튜플 개수이고, $R_{x1}, R_{x2}, R_{y1}, R_{y2}$이 현재 버켓의 부모 슬롯과 위상 관계 영역과의 겹침 비율이라고 할 때 공간 술어의 선택도는 다음과 같은 수식을 이용하여 산출한다.

$$ef \leftarrow f * R_{x1} * R_{x2} * R_{y1} * R_{y2},$$

Algorithm EContain은 이를 바탕으로 한 개의 질의 MBR이 주어졌을 경우 히스토그램을 사용하여 포함 관계 술어의 선택도를 추정하기 위한 알고리즘이다.

```
Algorithm EContain( N_x1 : node, x1, y1, x2, y2:int)
  Var R_x1, R_x2, R_y1, R_y2 : real
 S_x1 = FindMaxMin(N_x1, x1)
  If ( S_x2 <> nil) Then ' if found
    Do
      N_x2 = GetChild( S_x1 )
      S_x2 = FindMinMax( N_x2, x2 )
      If ( S_x2 <> nil) Then ' if found
        Do
          N_y1 = GetChild( S_x2 )
          S_y1 = FindMaxMin( N_y1, y1 )
          If ( S_y1 <> nil) Then
            Do
              N_y2 = GetChild( S_y1 )
              S_y2 = FindMinMax( N_y2, y2 )
              If ( S_y2 <> nil) Then
                Do
                  f = GetFreq( S_y2 )
                  R_x1 = GERatio( S_x1, x1 )
                  R_x2 = LERatio( S_x2, x2 )
                  R_y1 = GERatio( S_y1, y1 )
```

```
            R_y2 = LERatio( S_y2, y2 )
          ef = ef + f * R_x1 * R_x2 * R_y1 * R_y2
            S_y2 = GetLT( S_y2 )
         Loop While S_y2 <> nil
       End
        S_y1 = GetGT( S_y1 )
     Loop While S_y1 <> nil
   End
    S_x2 = GetLT( S_x2 )
 Loop While S_x2 <> nil
End
 S_x1 = GetGT( S_x1 )
Loop While S_x1 <> nil
End
End
```

4. 선택도 추정 기법의 성능 평가

실험은 다수의 공간 데이터 튜플들로부터 샘플링을 수행하여 구축된 공간 히스토그램을 이용하여 다수의 포함 질의의 선택도를 추정한 후 실제 공간 데이터 튜플에서 해당 질의를 수행한 결과와의 차이를 추정 오류 3 기법에 의하여 비교하였다. 실험에서 사용한 데이터는 질의의 대상이 되는 공간 데이터와 질의 술어를 위한 질의 MBR을 균일 분포, 정규 분포 등의 통계적인 방법을 사용하여 생성하였다.

4.1 성능 평가 환경

성능 평가는 모집단의 분포 특성, 샘플링 크기, 질의 MBR의 분포 특성, 분할 전략, 분할 개수 등을 다음과 같이 변경시키며 수행하였다.

모집단의 튜플 개수는 10,000개, 질의 MBR의 개수는 1,000개이다. 모집단 MBR의 위치 및 크기는 실세계에서 공간 객체의 분포에 따른 변화를 모델링 하기 위하여 균일 분포(uniform distribution) 및 정규 분포(normal distribution)에 따른 난수를 이용하였다. MBR의 중앙 점 분포는 전 영역에 균일하게 분포하는 경우(균일분포 : 최소 0, 최대 100)와 영역의 중앙에 집중적으로 분포하는 경우(정규 분포 : 평균 50, 분산 30)를 실험하였다. MBR의 크기 분포는 전영역에 분포하는 경우(균일분포 : 최소 0, 최대 100)와 작은 크기(정규분포: 평균 20,분산 30)와 큰 크기(정규분포 : 평균 80, 분산 30)를 사용하였다.

질의 MBR의 분포상 특징은 중앙 점 분포의 경우 전 영역에 균일하게 분포하는 경우(최소 0, 최대 100)만으로 한정하였으며 크기는 전영역에 분포하는 경우(균일분포 : 최소 0, 최대 100)와 작은 크기(정규분포 : 평균 20,분산 30)와 큰 크기(정규분포 : 평균 80, 분산 30)를 사용하였다.

히스토그램 구축을 위한 샘플링은 모집단 크기 10,000건에 대하여 신뢰도 95%를 만족하기 위하여 99(±10%), 385(±5%), 588(±4%), 1000(±3%), 2000(±2%), 5000(±1%)으로 변경하며 실험하였다.

실험에 사용된 분할 전략은 FindBoundary 함수의 변형을 이용하여 equi-width, equi-depth 두 가지를 사용하였다. 분할의 개수는 각 좌표축에 대하여 1에서 6개(버켓 개수= $1^4 \sim 6^4$)로 변경하며 실험하였다.

본 실험은 윈도우즈 NT 4.0상에서 데스크탑 지리정보 시스템인 GEOBase의 질의 처리 모듈을 확장하여 수행되었다.

4.2 질의 술어의 선택도 추정 오류

질의 술어의 선택도 추정 오류는 p-버켓에 기인한다. N개의 튜플을 가지고 있는 릴레이션에서 주어진 공간 위상 질의를 실제 만족하는 튜플의 개수를 N_q라 하고 히스토그램을 이용하여 추정한 튜플의 개수를 N_e라고 할 경우 선택도 추정시의 오류는 다음과 같이 다양한 방법으로 정의될 수 있다.

$$E = \frac{N_e}{N_q} \qquad \text{<추정 오류 1>}$$

$$E = \frac{|N_e - N_q|}{N_q} \qquad \text{<추정 오류 2>}$$

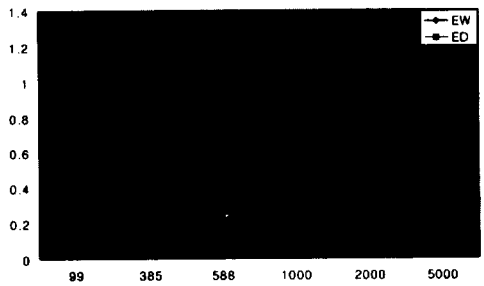
$$E = \frac{|N_e - N_q|}{N} \qquad \text{<추정 오류 3>}$$

추정 오류 1은 추정이 정확할 경우 1에 근접하고, 과대 및 과소 추정의 경우 각각 1이상, 1이하의 값을 가진다. 추정 오류 2는 추정이 정확할 경우 0에 근접하고 그렇지 않을 경우 0 이상의 값을 가진다. 이 두 기법은 N_q 및 N_e가 작은 값을 가질 경우 산정된 오류가 필요 이상 민감해 지는 문제를 가지고 있다. 즉, N_q=5이고, N_e=10일 경우와 N_q=500이고, N_e=1000일 경우 모두 각각 2.0와 1.0을 오류로 산정하지만 전체 튜플의 개수가 10000건이라고 가정할 경우 전자의 경우 질의 수행에 큰 차이를 보이지 않는 정도의 오류이며, 질의 최적화 단계가 필요 이상으로 민감해

지는 문제를 발생시킨다. 추정 오류 3은 이러한 문제를 해결하기 위하여 추정된 튜플 개수와 실제 튜플간의 차이를 전체 릴레이션의 크기에 대한 비율로 오류를 산정한다. 본 논문에서는 공간 술어의 추정 시 오류를 평가하기 위하여 추정 오류 3을 사용한다.

4.3 제안된 추정 기법의 성능 평가 결과

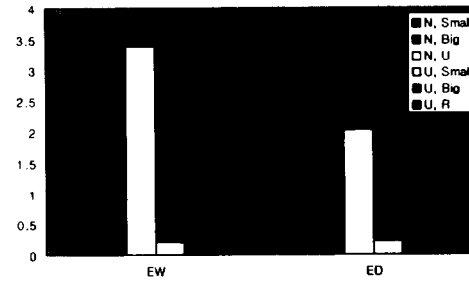
먼저 샘플 크기 변화에 따른 추정 오류의 변화를 실험하였다. 이 실험의 결과에 따라 모집단의 크기에 따른 적절한 샘플링 크기를 짐작할 수 있다. 이 실험에서는 모집단과 질의 MBR의 위치 및 크기를 균일 분포로, 분할의 개수를 각 축 당 6개로 고정 시키고 샘플링의 크기를 99개에서 5000개 사이로 변화 시키며 두 가지 분할 기법에 대한 추정 오류를 평가하였다. (그림 7)에 따르면 두 가지 분할 기법 모두 샘플링의 크기가 커짐에 따라 오류가 감소하고 있음을 볼 수 있으나 감소 폭은 크지 않음을 알 수 있다. 이는 95%의 신뢰도에 따른 샘플링 크기의 변화에 따른 정확도의 변화가 ±10%, ±5%, ±4%, ±3%, ±2%, ±1%로 미미하다는 것으로부터 유추할 수 있는 범위이다. 본 실험에 따르면 모집단의 개수에 대하여 $\frac{1}{20}$ 미만의 샘플링을 통한 히스토그램 구축으로도 상당히 정확한 수준의 공간 위상 술어의 선택도를 추정해 낼 수 있음을 의미한다. 샘플링의 크기가 커질 수록 히스토그램을 저장하기 위한 기억공간이 증가하지만 일정 크기 이상의 샘플링 크기에서는 저장 공간의 증가에 비해 추정의 정확도가 더 이상 증가하지 않음을 보여준다.



(그림 7) 샘플 크기 변화에 따른 추정 오류
(Fig. 7) Estimation error by sample size change

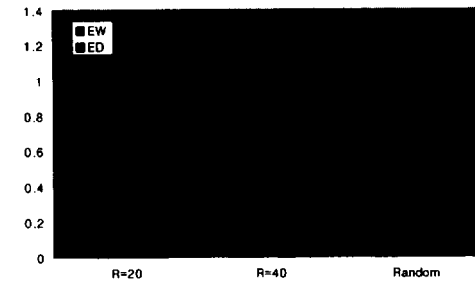
다음은 모집단의 분포 특성에 따른 오류 분석이다. 이 실험에서는 분할의 개수는 각 축 당 4개로, 샘플링 크기는 588개로 고정시킨 상태에서 모집단 MBR의 위치를 균일분포(U), 정규분포(N)로 MBR의 크기를 대(Big), 소(Small), 임의(R)로 변화시키며, 두 가지 분할 기법의 선택도 추정 오류의 변화를 실험하였다.

(그림 8)의 실험 결과에 따르면 두 가지 분할 기법 모두 모집단의 MBR이 균일 분포이고 MBR의 크기가 작을 경우 추정 오류가 감소됨을 볼 수 있다. 실세계의 지리정보 시스템 등에서는 MBR의 위치가 집단적으로 분포하는 경우가 많으며, 작은 크기의 MBR이 주를 이루는 경우가 많으며, 이러한 특성이 제안된 기법에 영향을 미칠 수 있음을 보인다.



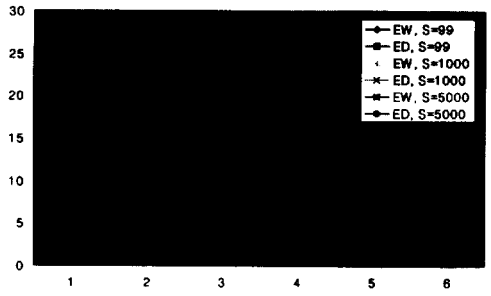
(그림 8) 모집단 분포 변화에 따른 추정 오류
(Fig. 8) Estimation error by data distribution change

다음 실험은 질의 MBR 크기에 따른 추정 오류의 변화에 대한 실험이다. 이 실험에서는 모집단의 분포는 위치 및 크기를 모두 균일 분포로, 분할의 개수는 각 축 당 4개, 샘플링의 크기는 1000으로 고정시키고 질의 MBR의 크기를 소, 대, 임의로 변경하며 실험하였다. (그림 9)의 실험 결과에 따르면 두 가지 기법 모두 MBR의 크기가 클 경우 추정 오류가 감소됨을 보이고 있다. 이는 MBR의 크기가 커짐에 따라 참조되는 p-버켓의 증가에 비해 f-버켓의 증가가 더 급격하기 때문으로 추정 오류의 감소 효과를 거둘 수 있다.



(그림 9) 질의 MBR 크기 변화에 따른 추정 오류
(Fig. 9) Estimation error by query MBR size change

다음은 분할 개수에 따른 추정 오류의 변화를 실험하였다. 이 실험은 모집단의 분포 및 질의 MBR의 분포를 균일 분포, 샘플링의 크기를 588로 고정시키고, 각 축 당 분할 개수를 1에서 6까지 변화시키며 수행하였다. (그림 10)은 이 결과로서 분할의 개수가 늘어남에 따라 추정 오류가 급격히 감소함을 보이고 있다. 이는 같은 질의 MBR에 의해 참조되어야 하는 p 버켓의 수에 비해 f-버켓의 수를 증가시키기 때문이며, 추정 오류의 원인이 p-버켓에 있으므로 상대적으로 추정 오류의 감소를 유도한다.



(그림 10) 분할 개수에 따른 오류
(Fig. 10) Estimation error by partition change

5. 결 론

본 논문에서는 기존의 공간 색인 기법으로 많이 사용되고 있는 차원 변환 기법이 공간 히스토그램을 이용한 추정 기법에서도 사용될 수 있음을 보이고 이를 이용하여 기존의 다차원 히스토그램을 이용한 추정 기법을 인접, 포함, 중첩 등의 공간 위상 술어의 선택도를 추정하기 위한 추정 기법을 제안하였다. 제안된 기법은 원 공간의 공간 객체들이 변환 공간상에서 점으로 대응된다는 점을 이용하여 기존의 분할 기법들을 변환 공간상에 적용 시켰으며 이를 이용한 히스토그램을 이용하여 다양한 공간 위상 관계의 선택도를 추정하기 위한 알고리즘이다. 제안된 기법은 기존의 속성 자료를 위한 히스토그램이 시스템 메모리 자원을 적게 사용하면서도 효과적인 술어 추정 기법으로 사용된다는 특징을 같이 가지고 있으며, 다양한 분할 전략을 채용하여 실세계 공간 객체의 다양한 분포 특성에 쉽게 대처할 수 있는 특징을 가진다.

본 기법이 비교적 작은 비용으로 공간 위상 술어의 선택도 추정에 사용될 수 있으나, 추정 기법이 참조해야 하는 노드의 개수가 많은 점, p-버켓 내의 오류를 최소화하기 위한 다양한 비율 산정 기법 등 다양한 분할 전략의 적용 등에 대한 추가적인 연구가 필요하다.

참 고 문 헌

[1] Blasgen, M. W., and Eswaran, K. P. "Storage and access in relational databases," IBM System Journal, 16(4), 1977.

[2] Selinger, P. G., et. al., "Access path selection in a relational database management system," Proc. of ACM SIGMOD, 1979.

[3] Yao, S. B., "Approximating block accesses in database organizations," CACM 20(4), pp.260-261, Apr., 1977.

[4] Shapiro, G. P., Connel, C., "Accurate estimation of the number of tuples satisfying a condition," ACM SIGMOD, pp.256-276, 1984.

[5] M.V. Mannino., P. Chu, and T. Sager, "Statistical profile estimation in database systems," ACM Computing Surveys, 20(3), pp.192-221, Sep., 1988,

[6] R. P. Kooi, "The optimization of queries in relational databases," Ph.D. thesis, Case Western Reserver University, Sept., 1980.

[7] R.J. Lipton, J.F. Naughton, and D.A. Schneider, "Practical selectivity estimation through adaptive sampling," Proc. of ACM SIGMOD Conf., pp.1-11, May, 1990.

[8] P. J. Haas and A. N. Swami, "Sampling-based selectivity estimation for joins using augmented frequent value statistics," Proc. of IEEE Conf. on Data Engineering, pp.522-531, 1995.

[9] C. M. Chen and N. Roussopoulos, "Adaptive selectivity estimation using query feedback," Proc. of ACM SIGMOD Conf, pp.161-172, May, 1994.

[10] W. Sun, Y. Ling, N. Rishe, and Y. Deng, "An instant and accurate size estimation method for joins and selections in a retrieval-intensive environment," Proc. of ACM SIGMOD Conf, pp. 79-88, 1993.

[11] V. Poosala, Y. E. Ioannidis, P. J. Haas, E. J.

Shekita, "Improved Histograms for Selectivity Estimation of Range Predicates," ACM SIGMOD Conf, pp.294-305, 1996.

[12] M. Muralikrishna, D. J. DeWitt, "Equi-Depth Histograms For Estimating Selectivity Factors For Multi-Dimensional Queries," ACM SIGMOD Conf, pp.28-36, 1988.

[13] V. Poosala, Histogram Based Estimation Techniques in Database Systems, Ph.D. Thesis, 1997.

[14] K, Y, Whang, S. W. Kim, and G. Wiederhold, "Dynamic Maintenance of Data Distribution for Selectivity Estimation," VLDB Journal, Vol.3, pp. 29-51, 1994

[15] Guttman, A., "R-trees : a dynamic index structure for spatial searching," Proc. ACM SIGMOD Conf., pp.47-57, 1984

[16] Nievergelt, J., Hinterberger, H., Sevcik, K.C., "The grid file : an adaptable, symmetric multikey file structure," ACM TODS, Vol.9, No.1, pp.38-71, 1984.

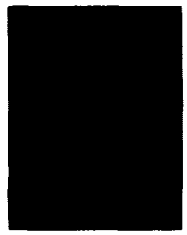
[17] B. Seeger, H. P. Kriegel, "Techniques for Design and Implementation of Efficient Spatial Access Methods," Proc. VLDB Conf., pp.360-371, 1988.

[18] Merrett, T. H., and Otoo, E., "Distribution models of relations," Proc. of VLDB, pp.418-425, 1979.

[19] Y. Ioannidis and S. Christodoulakis, "Optimal histograms for limiting worst-case error propagation in the size of join results," ACM TODS, 1993.

[20] Y. Ioannidis, "Universality of serial histograms," Proc. of VLDB, pp.256-267, Dec., 1993.

[21] Y. Ioannidis and V. Poosala, "Balancing histogram optimality and practicality for query result size estimation," Proc. of ACM SIGMOD Conf, pp.233-244, May, 1995.



김 홍 연

e-mail : redkite@netsgo.com

1992년 인하대학교 통계학과 학사

1994년 인하대학교 전자계산 공학과 석사

1995년~현재 인하대학교 전자계산공학과 박사과정 재학 중

관심분야 : 지리정보시스템, 공간데이터베이스, 저장관리자, WWW지리정보시스템

배 해 영

e-mail : hybae@dragon.inha.ac.kr

1974년 인하대학교 응용물리학과 (공학사)

1978년 연세대학교 대학원 전자계산학과(공학석사)

1989년 숭실대학교 대학원 전자계산학과(공학박사)

1985년 Univ. of Houston 객원 교수

1992년~1994년 인하대학교 전자계산소 소장

1982년~현재 인하대학교 전자계산 공학과 교수

관심분야 : 데이터베이스, 지리정보시스템, 리얼타임 데이터베이스