공간 데이터베이스 시스템을 위한 액세스 구조의 물리적 데이터베이스 설계 기법

이 종 학† · 박 병 권††

요 약

본 연구에서는 공간 데이터베이스 시스템에서 변환기법을 이용한 공간 액세스 구조에 대한 물리적 데이터베이스의 설계기법을 제안한다. 지금까지 공간 액세스 구조에 대한 많은 연구가 수행되어 왔으나 공간의 물리적 데이터베이스 설계의 측면에서의 연구는 수행된 바가 없다. 본 논문에서는 먼저 원공간(original space)에 주어지는 모든 공간 질의가 변환기법에 의한 변환공간에서는 한 가지 형태의 범위 질의로 변환되는 특징이 있음을 보인다. 그리고 변환공간상에서 이 범위 질의가 위치하는 질의 영역의 모양과 데이터 페이지가 위치하는 페이지 영역의 모양 사이의 관련성을 이용하여 질의처리의 성능을 향상시킬 수 있는 공간 액세스 구조의 최적 구성 기법을 제안한다. 또한, 성능평가를 위하여 공간 액세스 구조의 하나인 MBR-MLGF를 이용하여 다양한 질의 패턴과 데이터 분포에 대하여 제안된 설계기법을 적용한 실험 결과를 제시한다. 실험 결과에 의하면, 제안된 기법은 주어진 질의 패턴에 따라 최적의 MBR-MLGF를 구성할 수 있으며, 이차원 원공간에 대하여 변환공간으로 일반화한 사차원 질의 영역의 구간비가 1 : 16 : 256 : 4096인 경우, 기존의 순환 분할 기법에 비해 질의처리의 성능이 다섯배 이상으로 향상된다. 이러한 질의처리 성능의 향상은 제안된 물리적 데이터베이스 설계기법이 매우 유용함을 나타내는 것이다.

A Physical Database Design Method for Access Structures of Spatial Database Systems

Jong-Hak Lee† · Byung-Kwon Park††

ABSTRACT

This paper presents a physical database design methodology for spatial access structures using transformation techniques in spatial database systems. Recently, many spatial access structures have been proposed in the literature. However, there has been no effort for their physical database design. We first show that most spatial queries in the original space are transformed into one type of range queries in the transform space, and then propose a method for finding the optimal configuration of spatial access structures by using the relationship between the shapes of query regions, that are correspond to the range queries, and page regions, that are correspond to data pages, in the transform space. For performance evaluation, we perform extensive experiments with the MBR-MLGF, a spatial access structure using transformation techniques, using various types of queries and data distributions. The results indicate that our proposed method builds optimal MBR-MLGF according to the query types. When the interval ratio of a transformed four-dimensional query region is 1 : 16 : 256 : 4096, the performance of the proposed method is enhanced by as much as five times over that of the conventional cyclic splitting method. The result confirms that the proposed physical database design methodology is useful in a practical way.

키워드 : 공간 데이터베이스(spatial databases), 공간 액세스 구조(spatial access structure), 공간 질의(spatial query), 질의 처리(query processing)

1. 서 론

지리 정보 시스템(Geographic Information System : GIS), VLSI 설계 등의 새로운 데이터베이스 응용에서는 공간상에서 크기(extension)를 가지는 객체인 공간 객체(spatial objects, 명확히 구별될 경우에는 간략히 객체라 표현)의 효율적인 관리가 필요하다[1]. 최근들어 이를 위하여 공간 데이터베이스 시스템(spatial database systems)에 대한 연구가 활발히 진행되고 있으며, 그 중에서도 공간 객체에 대한 데이터 모델링과 액세스 방법 분야에 연구가 집중되고 있다[2].

공간 데이터를 위한 액세스 방법으로 공간상에서 인접한 객체들을 한꺼번에 검색하는 등의 공간 질의 고유의 특성을 고려하는 새로운 공간 액세스 방법(Spatial Access Method : SAM)이 필요하다. 공간 객체는 공간상에서 크기를 가지므로 SAM은 객체의 크기를 관리할 수 있는 메카니즘을 가져

※ 본 연구는 2000학년도 대구가톨릭대학교 연구비 지원에 의한 것임.
† 종신회원 : 대구가톨릭대학교 컴퓨터정보통신공학부 교수
†† 정 회 원 : 동아대학교 경영정보과학부 교수
논문접수 : 2001년 10월 8일, 심사완료 : 2002년 1월 4일

야 한다. 이러한 SAM에는 공간 객체의 크기를 관리하는 방법에 따라 공간 채움 곡선 기법(space filling curve techniques)[3-5], 객체 분할 또는 중복 기법(object clipping or object duplication techniques)[6], 영역 겹침 기법(region overlapping techniques)[7-9], 변환기법(transformation techniques)[10-12] 등으로 분류할 수 있다[13].

SAM의 한 부류인 변환기법(transformation techniques)은 원공간(original space)에서의 공간 객체의 특징을 대표할 수 있는 파라미터들을 사용하여 크기를 가지는 객체들을 변환공간(transform space)내의 점 객체들로 변환하여 관리함으로써 객체의 크기에 대한 별도의 관리를 없애기 위한 방법이다[10, 11]. 지금까지 연구된 대표적인 두 변환기법에는 구석점 변환(corner transformation)과 중앙점 변환(center transformation) 기법이 있다[10-12]. 이들 기법들은 포함자로서 객체를 최소한으로 포함하는 축에 수직인 선분들로 구성되는 최소 경계 사각형(Minimum Bounding Rectangle : MBR)을 이용하여 객체를 색인한다. 구석점 변환기법은 이차원 원공간에서 객체에 대한 MBR의 좌하(lower-left)점과 우상(upper-right)점의 네 좌표 값을 파라미터로 사용하고, 중앙점 변환기법은 객체에 대한 MBR의 중심점의 두 좌표값과 각 변 길이의 반을 네 파라미터로 사용하여 객체를 사차원 변환공간내의 한 점으로 표현한다. 변환기법과 같은 공간 액세스 방법이 공간 데이터베이스 시스템에서 실제로 많이 사용되기 위해서는 변환기법의 특성에 대한 연구와 이를 이용한 공간 질의의 기본 유형들에 대한 효율적인 질의처리 방법 등이 연구되어야 한다[12].

일반적으로 데이터베이스 관리 시스템(database management system : DBMS)의 성능을 평가하는 기준의 하나는 효율적인 질의처리를 통하여 사용자 질의에 대하여 얼마나 빠른 응답을 할 수 있는가 하는 것이다. 질의처리는 전체 레코드들 중에서 질의 조건을 만족하는 것만을 찾아내는 과정이다. DBMS의 성능 향상을 위한 물리적 데이터베이스 설계(physical database design)[14-17]는 사전에 분석한 질의 정보를 기반으로 데이터베이스 파일의 클러스터링 특성[18]과 액세스 방식 등을 결정함으로써 주어진 데이터베이스 응용의 질의처리 시 최적의 성능을 지원하는 방법이다[17]. 따라서 공간 데이터베이스 시스템을 위한 액세스 구조에서도 이러한 물리적 데이터베이스 설계가 필요하다. 본 논문에서는 변환기법을 이용한 공간 액세스 구조에 대하여 데이터베이스의 클러스터링 특성과 관련된 물리적 데이터베이스 설계기법을 제안하고자 한다.

본 논문에서는 일련의 공간 데이터베이스 사용자 질의 패턴에 대하여 평균 질의처리 비용을 최소화할 수 있는 공간 액세스 구조를 구성하기 위하여 도메인 공간(domain space)[19]의 분할전략(splitting strategy)을 결정하는 문제를 *공간 액세스 구조를 위한 물리적 데이터베이스 설계(Physical Database Design for Spatial Access Structures)*라 정의한다. 지금까지 제안된 공간 액세스 구조에서는 각 축을 미리 결정된 일정한 순서에 의해서 번갈아 가면서 분할한다. 이러한 방식은 사용자로부터 주어지는 질의 패턴의 특성을 전혀 반영하지 않은 것이다.

일반적으로, 사용자로부터 주어지는 공간 데이터베이스 사용자 질의 패턴에서 공간 액세스 구조를 구성하는 각 축에 대한 질의 조건이 동일한 크기의 범위로 주어지는 경우는 거의 없다. 또한, 서로 다른 크기의 도메인을 사용하는 변환공간의 두 축에 대해서는 질의 조건이 동일한 크기의 범위로 주어지는 경우에도 변환된 질의 영역에 대한 두 축의 구간 크기는 매우 달라진다. 예를들어, 지리 정보 시스템의 응용을 위한 공간 액세스 구조의 하나인 MBR-계층 그리드 파일(MBR-Multilevel Grid File : MBR-MLGF)[2]에서 X축은 $100KM$의 도메인이 할당되고, Y축은 $10KM$의 도메인이 할당되었을 때, $100M \times 100M$의 범위 질의는 변환공간에서 $1:10$ 비율의 질의 영역으로 변하게 된다.

변환공간상의 범위 질의가 위치하는 영역을 *질의 영역(query region)*이라 할 때, 변환기법을 이용한 공간 액세스 구조에서 원공간에 주어지는 모든 공간 질의가 변환공간에서는 한 가지 형태의 범위 질의로 변환되므로, 공간 데이터베이스 시스템의 사용자 질의는 변환공간내의 질의 영역에 포함되는 객체들을 탐색하는 연산으로 해석할 수 있다. 따라서, 공간 액세스 구조의 데이터 페이지에 해당하는 변환공간상의 영역을 *페이지 영역(page region)*이라 할 때, 공간 액세스 구조의 물리적 데이터베이스 설계의 문제는 일련의 사용자 질의 영역들에 의해서 교차되는 페이지 영역들의 개수를 최소화하는 문제가 된다.

본 논문에서는 영역을 구성하는 각 축의 구간 크기에 대한 비율을 *구간비(interval ratio)*라 정의하고, 영역의 모양을 구간비로서 표현한다. 공간 사용자 질의 패턴에 나타나는 질의 영역들의 모양에 대한 정보를 이용하여 사용자 질의의 질의 영역들에 의해 교차되는 페이지 영역들의 개수가 최소로 되는 최적의 구간비를 결정하고, 공간 액세스 구조의 도메인 공간이 가능한 이와 같은 구간비를 갖는 페이지 영역들로 구성되도록 하는 영역 분할전략을 사용함으로써 최적의 공간 액세스 구조를 구성할 수 있다.

본 논문의 구성은 다음과 같다. 제 2절에서는 관련 연구로서 물리적 데이터베이스를 적용할 대상으로 변환기법을 이용한 공간 액세스 구조를 소개한다. 제 3절에서는 먼저 공간 데이터베이스 시스템에 주어지는 모든 공간 질의는 한 가지 형태의 범위 질의로 변환되는 특징에 대하여 기술하고, 이러한 특징을 이용하여 공간 질의들을 가장 효율적으로 처리할 수 있는 공간 액세스 구조의 물리적 데이터베이스 설계 기법을 제안한다. 그리고, 제 4절에서는 성능 평가를 위한 실험 환경과 실험 결과를 제시한다. 마지막으로, 제 5절에서는 결론을

내린다.

2. 관련 연구

본 절에서는 공간 액세스 구조의 물리적 데이터베이스 설계를 적용하기 위한 공간 액세스 구조로서 구석점 변환기법을 이용한 공간 액세스 구조인 MBR-MLGF[2]을 소개한다. 먼저, MBR-MLGF의 기본 구조인 계층 그리드 파일(Multilevel Grid File : MLGF)[19, 20]의 특성에 대하여 설명한 다음에 MBR-MLGF의 정의와 그 구조적 특성과 장점에 대하여 설명한다.

2.1 MLGF

MLGF[19]는 디렉토리와 데이터 페이지들로 구성된다. 디렉토리는 균형 트리 구조를 가지며, 각 단계 디렉토리는 전체 공간의 분할 상태를 반영한다. 디렉토리의 최하위 단계에 있는 엔트리(디렉토리 엔트리)는 데이터 페이지를 가리킬 뿐만 아니라, 그 페이지에 할당된 영역(페이지 영역)을 표현한다. 하나의 데이터 페이지는 디렉토리 엔트리에 의해서 표현된 영역내에 속하는 데이터 레코드들만을 저장한다. 그리고 다단계의 디렉토리 구조는 재귀적으로 구성된다. 즉, 상위 단계의 디렉토리 엔트리는 차하위 단계의 디렉토리 페이지를 가리키며, 그 디렉토리 페이지가 가리키는 영역을 표현한다.

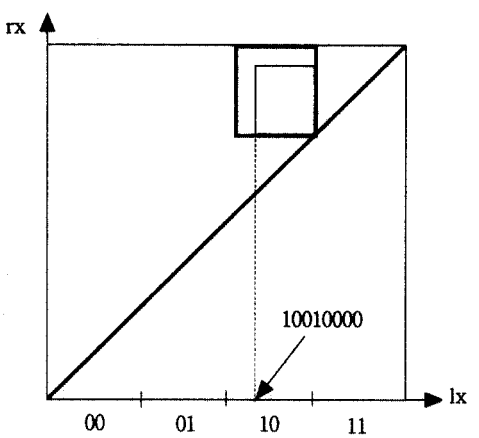
MLGF의 디렉토리 엔트리는 영역벡터(region vector)와 다음 단계 페이지에 대한 포인터로 구성된다. N개의 속성을 갖는 MLGF의 영역벡터는 N개의 해쉬값으로 구성되며, 해당 디렉토리 엔트리가 나타내는 영역의 위치, 크기, 모양에 대한 정보를 갖는다. 영역벡터에서 i번째 해쉬값은 그 디렉토리 엔트리의 영역내에 속하는 모든 객체들의 i번째 속성을 해싱하였을 때 나타나는 해쉬값들의 공통 접두부(common prefix)가 된다. 또한 상위 단계의 한 영역은 그 영역에 대응하는 엔트리가 가리키는 부트리(subtree) 내의 모든 영역을 포함한다.

MLGF는 객체가 공간에 삽입되고 삭제되는 상황에 따라 분할과 병합을 반복함으로써 동적 변화에 적응한다. 디렉토리 또는 데이터 페이지에 오버플로우가 일어날 때 분할되는 영역은 구간 이등분 정책을 사용하여 같은 크기의 두 영역으로 분할된다. MLGF에서 공간을 분할할 때 나타나는 가장 큰 특징은 분할이 요구되는 영역만을 분할시키는 부분 분할 방식(local splitting strategy)[20]을 취한다는 것이다. 이러한 분할 방식은 반드시 필요한 영역만을 생성시켜 디렉토리 엔트리 수의 증가를 억제한다. 이 결과 MLGF의 디렉토리는 저장되는 객체의 분포나 서로 다른 속성간의 상관 관계 등 데이터 특성에 큰 영향을 받지 않고 삽입되는 객체 수에 선형적으로 비례하여 증가한다[19, 20]. 따라서 MLGF는 구석점 변환기법을 사용하는 경우 발생하는 상관 관계가 높은 비균일 데이터 분포에 잘 적응할 수 있다.

2.2 MBR-MLGF

MBR-MLGF[2]는 공간 객체를 효율적으로 관리할 수 있도록 MLGF를 확장하여 저장 공간의 오버헤드 없이 디렉토리 엔트리에 객체의 MBR을 추가로 유지하고 있다. 즉, MBR-MLGF의 각 디렉토리 엔트리에는 원공간에서의 객체에 대한 MBR을 표현하기 위하여 그 디렉토리 엔트리에 대응하는 영역 내에 존재하는 객체들의 최소 좌측점(*min-lx*) 값과 최대 우측점(*max-rx*) 값을 유지하고 있다.

MBR-MLGF 내의 한 영역에 대한 영역벡터 중에서 lx축에 대한 해쉬값은 그 영역 내에 위치하는 객체들의 lx값의 접두부(prefix)가 된다. 이 값은 또한 이 영역에 대한 극단값의 접두부가 된다. (그림 1) (a)는 MBR-MLGF의 한 디렉토리 엔트리의 대응 영역과 극단값(그림의 가는 선)을 표시한 것이다. 이 예에서 MBR-MLGF의 각 축의 해쉬값을 위한 최대 정밀도는 8비트로 표현된다고 가정한다. 이 그림에서 이 영역에 속하는 객체들에 대한 min-lx는 '10010000'이고 영역벡터의 lx축 해쉬값은 '10'이므로, '10'은 '10010000'의 접두부가 된다.



(a) 변환공간의 한 영역

2	1	0	×	×	×	×	×	×

(b) MLGF에서의 lx축 해쉬값의 표현

2	1	0	0	1	0	0	0	0

(c) MBR-MLGF에서의 lx축 해쉬값과 극단값의 동시 표현

(그림 1) MBR-MLGF에서의 극단값 표현

MLGF에서는 한 영역에 대한 영역벡터 내의 i번째 축에 대한 해쉬값을 표현하기 위해서 이 영역의 i번째 속성들의 공통 접두부와 그 길이를 엔트리 내에 기록한다. (그림 1) (b)는 MLGF의 디렉토리 엔트리에서 lx축에 대한 해쉬값을 표현하는 자료구조를 나타낸 것이다. 이 자료구조 앞부분의 숫

자는 공통 접두부의 길이를 나타내며, 뒷부분은 해쉬값을 비트 단위로 나타낸 것이다. 디렉토리 엔트리의 lx축에 대한 해쉬값 '10'은 이 자료구조 뒷부분에 표시되며, 이 해쉬값은 두 비트로 구성되므로 앞 부분에 2가 표시된다. 나머지 'x'로 표현된 여섯 비트는 이 엔트리를 위해서는 현재 사용되지 않는 부분이다.

MBR-MLGF의 각 디렉토리 엔트리에서 극단값은 MLGF의 각 축에 대한 해쉬값 구조에서 사용되지 않는 부분을 사용하여 표현할 수 있다. 따라서, 극단값 정보를 추가하는 경우에도 디렉토리 엔트리의 크기에는 변화가 없다. (그림 1) (c)에서 lx축에 대한 극단값 '10010000'은 8비트 모두를 사용하여 표현할 수 있다. 이중 앞 두 비트는 영역벡터의 lx축에 대한 해쉬값을 그대로 표현한다고 할 수 있다. 따라서 MBR-MLGF는 MLGF의 디렉토리 엔트리와 동일한 크기의 자료구조로 별도의 저장공간 오버헤드 없이 극단값을 추가로 유지할 수 있다.

3. 공간 액세스 구조의 물리적 데이터베이스 설계

본 절에서는 변환기법을 이용한 공간 액세스 구조의 물리적 데이터베이스 설계에 대하여 논한다. 먼저, 제 3.1절에서 변환기법을 이용한 공간 액세스 구조에서 원공간에 주어지는 모든 공간 질의가 변환공간에서는 한 가지 형태의 범위 질의로 변환되는 특징을 기술한다. 제 3.2절에서는 이차원 변환공간에 대해서 질의처리 시 발생하는 평균 페이지 접근 횟수를 최소로 하는 액세스 구조의 최적 조건과 함께, 이 조건을 만족하는 공간 액세스 구조의 영역 분할 전략을 제시한다. 그리고, 제 3.3절에서는 공간 데이터베이스의 가장 일반적인 사차원 변환구조에 대하여 물리적 데이터베이스 설계 알고리즘을 기술한다.

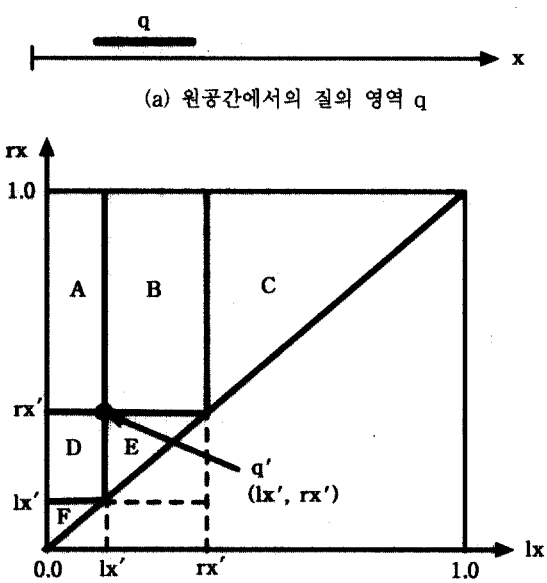
3.1 공간 질의의 특징

*공간 질의(spatial queries)*는 주어진 질의 영역에 대하여 특정 공간 관계를 가지는 객체들을 검색하는 질의이다. 공간 데이터 처리를 위한 요구조건을 모두 만족하는 공간 질의의 표준 집합은 아직 정해지지 않았으나, 공간 질의를 위한 기본 유형들은 여러 연구들[11, 21]에서 제안되었으며 이들을 종합하면 다음과 같이 정리할 수 있다.

- 영역 교차 질의(region intersection query) : 주어진 질의 영역과 교차하는 모든 공간 객체를 검색한다.
- 영역 포함 질의(region containment query) : 주어진 질의 영역 속에 포함되는 모든 공간 객체를 검색한다.
- 영역 피포함 질의(region enclosure query) : 주어진 질의 영역을 포함하는 모든 공간 객체를 검색한다.
- 점 질의(point query) : 주어진 질의 점을 포함하는 모든 공간 객체를 검색한다. 이는 영역 피포함 질의의 특수한 경우로 볼 수 있다.
- 최근접 인접 질의(nearest neighbor query) : 주어진 질의 점에서 가장 가까운 객체를 검색한다.

이들 기본 유형 중에서 영역 교차 질의, 영역 포함 질의, 영역 피포함 질의, 점 질의는 주어진 질의 영역(점 질의의 경우는 질의점)에 대한 특정 공간 관계를 만족하는 객체를 모두 검색하는 질의이므로, 이들을 통칭하여 *영역 질의(region query)*라 정의한다. 기본 유형의 공간 질의로 처리할 수 없는 질의가 요구될 때는 이들 질의들을 순차적으로 수행하여 원하는 질의를 처리할 수 있다.

본 절에서는 변환기법을 이용한 공간 액세스 구조에서 원공간에 주어지는 모든 공간 질의가 변환공간에서는 한 가지 형태의 *범위 질의(range query)*로 변환되는 특징이 있음을 알아본다. 먼저, 영역 질의의 처리를 위해서는 주어진 질의 영역과 특정 공간 관계를 가지는 객체들이 변환공간의 어느 부분에 위치하는가를 알아보자. (그림 2)는 일차원 원공간의 한 질의 영역 q를 이차원 변환공간상에 나타낸 질의 점 q'과 이를 중심으로 변환공간상에서 상대적 공간 관계에 따라 객체가 존재할 수 있는 영역을 A에서 F까지 여섯 영역으로 구분하여 나타낸 것으로 각 영역의 특징은 다음과 같다.



(a) 원공간에서의 질의 영역 q

A : 원공간에서 q을 포함하는 객체들이 존재하는 영역
B : 원공간에서 q의 오른쪽점과 교차하는 객체들이 존재하는 영역
C : 원공간에서 q의 오른편에 존재하는 객체들이 존재하는 영역
D : 원공간에서 q의 왼쪽점과 교차하는 객체들이 존재하는 영역
E : 원공간에서 q에 포함되는 객체들이 존재하는 영역
F : 원공간에서 q의 왼편에 존재하는 객체들이 존재하는 영역

(b) 변환공간에서의 공간 관계별 영역 구분

(그림 2) 변환기법의 공간 관계별 영역 구분

- 영역 A : 원공간에서 q를 포함하는 모든 객체들이 위치한다. 영역 A내의 모든 객체들의 lx값은 lx'(q'의 lx값)보다 작고 rx값은 rx'(q'의 rx값)보다 크기 때문이다.
- 영역 B : 원공간에서 q의 우측점과 교차하는 모든 객체들이 위치한다. 영역 B내의 모든 객체들의 lx값은 lx'과 rx' 사이이고 rx값은 rx'보다 크기 때문이다.
- 영역 C : 원공간에서 q의 우측에 존재하는 모든 객체들이 위치한다. 영역 C내의 모든 객체들의 lx값이 rx'보다 크기 때문이다.
- 영역 D : 원공간에서 q의 좌측점과 교차하는 모든 객체들이 위치한다. 영역 D내의 모든 객체들의 lx값은 lx'보다 작으며 rx값은 lx'과 rx' 사이이기 때문이다.
- 영역 E : 원공간에서 q에 포함되는 모든 객체들이 위치한다. 영역 E내의 모든 객체들의 lx값은 lx'보다 크고 rx값은 rx'보다 작기 때문이다.
- 영역 F : 원공간에서 q의 좌측에 존재하는 모든 객체들이 위치한다. 영역 F내의 모든 객체들의 rx값이 rx'보다 작기 때문이다.

위와 같은 영역 구분을 이용하면 원공간의 다양한 영역 질의는 변환공간에서의 범위 질의로 변환하여 처리할 수 있다. 다음은 주어진 질의 영역 q에 대한 영역 질의들이 어떤 범위 질의로 변환되는가를 나타낸다.

- 영역 교차 질의 : (그림 2) (b)의 변환공간에서 검색되어야 하는 부분은 A, B, D와 E를 합한 영역이다. 따라서, 이 영역은 이차원공간상에서 '$lx \le rx'$ AND $rx \ge rx'$'인 범위 질의의 처리를 위하여 검색해야 하는 부분이다.
- 영역 포함 질의 : (그림 2) (b)의 변환공간에서 검색되어야 하는 부분은 영역 E이다. 따라서, 이 영역은 이차원공간상에서 '$lx' \le lx \le rx'$ AND $lx' \le rx \le rx'$'인 범위 질의의 처리를 위하여 검색해야 하는 부분이다.
- 영역 피포함 질의 : (그림 2) (b)의 변환공간에서 검색되어야 하는 부분은 영역 A이다. 따라서, 이 영역은 이차원공간상에서 '$lx \le lx'$ AND $rx \ge rx'$'인 범위 질의의 처리를 위하여 검색해야 하는 부분이다.
- 점 질의 : 주어진 질의 점 q_p의 좌표를 $q_p.x$라고 할 경우 이 질의를 위하여 검색할 부분은 이차원 변환공간상에서 '$lx \le q_p.x$' AND $rx \ge q_p.x$'인 범위 질의의 처리를 위하여 검색해야 하는 부분이다.

3.2 공간 액세스 구조의 영역 분할전략

본 절에서는 먼저 변환공간상에서 범위 질의가 위치하는 질의 영역의 구간비와 액세스 구조의 데이터 페이지가 위치하는 페이지 영역의 구간비 사이의 상관관계로서 질의처리의 최적조건에 대하여 알아본다. 그리고, 이차원의 MBR-MLGF에 대해서 페이지 영역의 구간비가 질의처리의 최적 조건에 합당한 페이지 영역의 최적 구간비에 근접하도록 하는 영역 분할전략을 제시하고 이를 사차원 MBR-MLGF로 확장한다.

다차원 점 액세스 방법을 지원하는 다차원 파일구조에서는 액세스 구조를 구성하는 페이지 영역의 모양에 따라 주어진 질의 영역에 의해서 교차되는 페이지 영역의 평균 개수가 달라지는 특징이 있다. 참고문헌[22]에서는 이러한 특징을 이용하여 다차원 공간내의 데이터의 균일분포와 비균일분포 각각에 대하여 주어진 질의 영역들에 대해 페이지 영역의 평균 접근 횟수를 최소로 하는 페이지 영역의 최적 구간비를 계산하는 방법을 제안하였다. 본 절에서는 이를 소개하고, 공간 액세스 구조에 대한 페이지 영역의 최적 구간비를 이와같은 방법으로 계산한다.

먼저, 이차원 공간상에서 데이터가 균일하게 분포할 때 페이지 영역의 최적 구간비를 계산하는 방법에 대하여 요약하면 다음과 같다. 데이터가 균일하게 분포하면 도메인 공간을 구성하는 페이지 영역들의 크기가 일정하게 되며, 주어진 질의 영역들에 의해 교차되는 페이지 영역들의 개수를 최소로 하는 페이지 영역의 최적 구간비는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산할 수 있다. 즉, 크기가 $p(x) \times p(y)$로 일정한 페이지 영역들로 나누어져 있는 이차원 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i(y)$ $(i = 1, \cdots, n)$에 대해 각 질의 영역과 교차하게 되는 페이지 영역의 총 개수를 최소로 하는 최적 페이지 영역의 구간비($p(x) : p(y)$)는 $\sum_{i=1}^{n} q_i(x) : \sum_{i=1}^{n} q_i(y)$이다[22].

그리고, 이차원 공간상에서 데이터가 비균일하게 분포할 때 페이지 영역의 최적 구간비를 계산하는 방법에 대하여 요약하면 다음과 같다. 이차원 공간상에서 데이터가 비균일하게 분포한다는 것은 도메인 공간내의 위치에 따라 객체의 밀집도가 다름으로 인하여 페이지 영역의 크기가 위치에 따라 달라짐을 의미한다. 즉, 밀집도가 높은 곳은 밀집도가 낮은 곳에 비하여 많은 페이지가 할당되므로 각 페이지 영역의 크기는 작아지게 된다. 따라서, 비균일 분포의 경우에는 질의 영역에 의해 교차되는 페이지 영역의 개수가 질의 영역의 크기뿐만 아니라 질의 영역이 주어진 위치의 데이터 밀집도에도 비례하게 되므로, 균일분포에서와 같이 페이지 영역의 최적 구간비를 모든 질의 영역의 각 축별로 구간 크기를 단순히 더한 값의 비율로서 구할 수 없다. 이와 같은 경우에는 각 질의 영역의 크기에 대해 위치에 따른 데이터 밀집도를 가중치(weight)로 곱한 질의 영역의 형태를 *정규화된 질의 영역(normalized query region)*이라 하고, 이러한 질의 영역의 정규화를 통하여 페이지 영역의 최적 구간비를 계산할 수 있다. 즉, 서로 다른 크기의 페이지 영역들로 나누어져 있는 이차원 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i(y)$ $(i = 1, \cdots, n)$에 대해 각 질의 영역의 객체 밀집도를 $d_i (= \frac{n\_o}{q(x) \times q(y)}$, 단, n_o는 질의 영역 내의 객체 수이다.)라 할

때, 각 질의 영역과 교차하게 되는 페이지 영역의 총 개수를 최소로 하는 페이지 영역의 최적 구간비($p(x) : p(y)$)는 $\sum_{i=1}^{n} q_i(x)\sqrt{d_i} : \sum_{i=1}^{n} q_i(y)\sqrt{d_i}$이다[22].

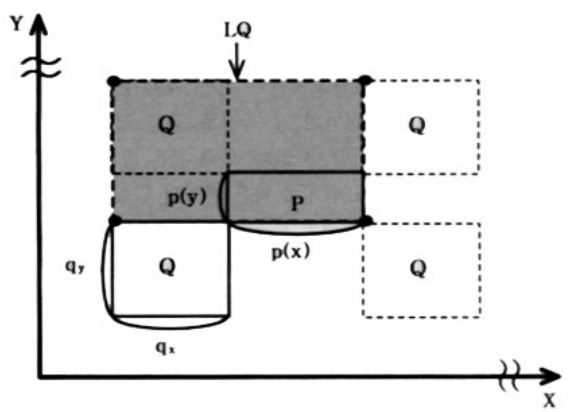
따라서, 본 논문에서는 먼저, 질의 패턴으로 주어지는 공간 질의들에 의해 다차원 변환공간상에 표현되는 질의 영역들로서 페이지 영역의 최적 구간비를 계산한다. 그리고, 공간 액세스 구조의 도메인 공간이 최적 구간비의 페이지 영역들로서 구성될 수 있도록 하는 영역 분할전략을 개발한다.

MBR-MLGF에서는 새로운 객체의 삽입으로 페이지의 용량이 초과되면, 이 페이지에 대응하는 페이지 영역은 같은 크기를 갖는 두 개의 영역으로 분할된다. 이때 분할되는 페이지 영역의 분할 축으로서 분할된 영역의 구간비가 최적 구간비에 가깝게 되는 축을 선택함으로써, 객체의 지속적인 삽입으로 인한 연속된 분할시에 도메인 공간내의 모든 페이지 영역의 구간비를 최적 구간비에 근접하도록 유도할 수 있다.

아래 정리 1은 이차원 공간상에 임의의 위치에 주어지는 특정 모양의 한 질의 영역이 특정 크기의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 범위의 크기는, 그 페이지 영역의 모양이 주어진 특정 질의 영역의 모양과 같을 때 최소가 됨을 나타낸다.

정리 1 구간비가 $q_x : q_y$인 $q_x \times q_y$형태의 질의 영역이 이차원 공간상에서 임의의 위치에 주어질 때, 크기가 B인 $p(x) \times p(y)$ 형태의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 범위의 크기는 페이지 영역의 구간비가 주어진 질의 영역의 구간비와 같을 때 최소가 된다.

증명 : 아래 (그림 3)은 이차원 도메인 공간에서 $q_x \times q_y$ 형태의 질의 영역 Q가 임의의 위치에 주어질 때, 크기가 $B(=p(x) \times p(y))$인 특정 페이지 영역 P와 교차하게 되는 위치의 범위를 질의 영역 Q의 좌상점이 위치할 수 있는 영역(빗금친 부분) LQ로 나타낸 것이다.



(그림 3) 임의의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 영역

(그림 3)에서 LQ의 크기 $SIZE_LQ(p(x), p(y))$는 다음 식과 같다.

$$SIZE\_LQ(p(x), p(y)) = (p(x) + q_x)(p(y) + q_y) \quad (1)$$

$p(x) \times p(y) = B$에 의해서, 식 (1)의 $p(y)$를 $\frac{B}{p(x)}$로 치환하면,

$$SIZE\_LQ\left(p(x), \frac{B}{p(x)}\right) = (p(x) + q_x)\left(\frac{B}{p(x)} + q_y\right) \quad (2)$$
$$= B + \frac{q_x B}{p(x)} + p(x)\,q_y + q(x)\,q_y$$

따라서, 식 (2)의 값을 최소로 하는 $p(x)$를 구하면, $p(x) = \sqrt{(q_x/q_y)B}$이다. 또한, 이러한 $p(x)$에 대한 $p(y)$는 $p(x) \times p(y) = B$에 의하여 $p(y) = \sqrt{(q_y/q_x)B}$이다. 그러므로, $SIZE_LQ(p(x), p(y))$를 최소로 하는 페이지 영역 P의 구간비 $p(x) : p(y) = q_x : q_y$이다. □

정리 1을 이용하여 페이지 영역의 분할 시 분할된 페이지 영역의 구간비가 최적 구간비에 가깝게 되는 분할 축을 선택할 수 있다. (그림 4)는 주어진 최적 구간비 $(a : b)$와 같은 모양을 갖는 $a \times b$형태의 질의 영역 Q가 이차원 도메인 공간상에 임의의 위치에 주어졌다고 가정하고, $p(x) \times p(y)$형태의 페이지 영역 P가 두개의 페이지 영역으로 분할된 후의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 범위(빗금친 부분)인 LQ를 나타낸다. (그림 4)에서 (a)는 분할 축으로 X축을 선택한 경우의 LQ_x를 나타내고, (b)는 분할 축으로 Y축을 선택한 경우의 LQ_y를 나타낸다.

(그림 4)에서 X축을 분할했을 경우의 LQ_x의 크기는

$$SIZE(LQ_x) = (p(x)/2 + a)(p(y) + b) \quad (3)$$

이고, Y축을 분할했을 경우의 LQ_y의 크기는

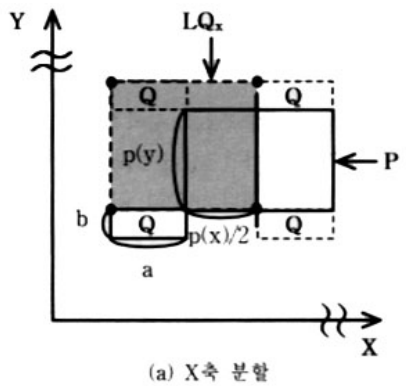
$$SIZE(LQ_y) = (p(x) + a)(p(y)/2 + b) \quad (4)$$

이다.

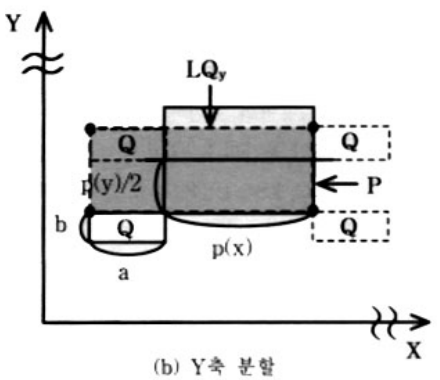
정리 1에 의하여 LQ의 크기는 페이지 영역의 구간비가 질의 영역의 구간비와 같을 때 최소가 되므로, 이 LQ의 크기가 작을수록 페이지 영역의 구간비가 질의 영역 Q의 구간비에 근접하게 된다. 따라서, 이 LQ의 크기가 작게 되는 축을 분할축으로 선택함으로써 분할 후의 페이지 영역의 구간비를 주어진 최적 구간비에 더 근접하게 할 수 있다.

MBR-MLGF에서는 페이지 영역의 각 축의 구간 크기는 디렉토리의 최하위 단계 디렉토리 엔트리의 영역벡터내에 유지된다((그림 1) 참조). 즉, 영역벡터의 한 축 X에 대한 해쉬값의 길이를 $l(x)$라 할 때, X축의 구간 크기는 $2^{l(x)}$에 반비

례한다. 따라서, $A(x)$를 X축 전체의 크기라 하면, 해당 페이지 영역의 X축의 크기($p(x)$)는 $A(x)/2^{l(x)}$이다. 그러므로, X축과 Y축으로 구성되는 이차원 MBR-MLGF에서 페이지 영역의 구간비를 $a:b$에 근접하게 하는 영역 분할전략은 식 (3)과 식 (4)에서 $p(x)$, $p(y)$에 $A(x)/2^{l(x)}$, $A(y)/2^{l(y)}$를 대입하여 정리하면, 만약 $(A(y)\times 2^{l(x)}\times a < A(x)\times 2^{l(y)}\times b)$이면 X축을 분할하고, 그렇지 않으면 Y축을 분할하면 된다.



(a) X축 분할



(b) Y축 분할

(그림 4) 분할후의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 영역

이차원 MBR-MLGF의 영역 분할전략을 사차원 MBR-MLGF에도 확장하여 적용할 수 있다. 즉, W, X, Y, Z축으로 이루진 사차원 MBR-MLGF의 페이지 영역의 분할 시 분할된 페이지 영역의 구간비가 주어진 최적 구간비에 가장 가깝게 되는 분할 축을 선택할 수 있다. 먼저, 최적 구간비($a:b:c:d$)와 같은 모양의 질의 영역 $Q(a\times b\times c\times d)$가 사차원공간상에 임의의 위치에 주어진다고 가정하고, 분할이 요구되는 페이지 영역($p(w)\times p(x)\times p(y)\times p(z)$)이 각 축에 대해 분할된 후의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 범위 LQ의 크기를 계산하여, 그 값이 가장 작게 되는 축을 분할 축으로 선택하면 된다. 예를 들어, 분할 축으로 W축을 선택했을 때 LQ의 크기($SIZE(LQ_w)$)는 $(p(w)/2+a)(p(x)+b)(p(y)+c)(p(z)+d)$이다. 따라서, 사차원 MBR-MLGF의 영역 분할 전략은 다음과 같다.

사차원 MBR-MLGF를 위한 영역 분할전략 :

- $(p(w)/2+a)(p(x)+b)(p(y)+c)(p(z)+d)$의 값이 최소이면, W축 분할,
- $(p(w)+a)(p(x)/2+b)(p(y)+c)(p(z)+d)$의 값이 최소이면, X축 분할,
- $(p(w)+a)(p(x)+b)(p(y)/2+c)(p(z)+d)$의 값이 최소이면, Y축 분할,
- $(p(w)+a)(p(x)+b)(p(y)+c)(p(z)/2+d)$의 값이 최소이면, Z축 분할.

3.3 공간 액세스 구조의 구축 알고리즘

(그림 5)는 제 3.2절에서의 이차원 변환공간에 대한 질의 영역과 페이지 영역간의 상호관계와 영역 분할전략을 사차원으로 확장하여 이차원 원공간에 대해 일반화한 공간 액세스 구조의 물리적 데이터베이스 설계 알고리즘을 나타낸다.

(그림 5)의 알고리즘에서 나타낸 공간 액세스 구조의 전체 설계 과정은 다음과 같은 세 가지 단계로 구성된다. 첫째, 이차원 원공간의 질의 패턴을 구성하는 각 질의에 대하여 사차원 변환공간상의 사차원 질의 영역으로 변환하여 정규화 과정을 거친다. 즉, 사차원 변환공간상의 사차원 질의 영역 $q(w)\times q(x)\times q(y)\times q(z)$에 대한 정규화는 다음과 같다. 먼저, 질의 결과에 의한 질의 영역내의 객체의 개수 n_o를 이용하여 레코드 밀집도 d를 $\dfrac{n_o}{q(w)\times q(x)\times q(y)\times q(z)}$로 구하고, 질의 영역을 이루는 각 축의 구간에 가중치 $d^{1/4}$을 곱하여 정규화된 질의 영역 $q(w)\,d^{1/4}\times q(x)\,d^{1/4}\times q(y)\,d^{1/4}\times q(z)\,d^{1/4}$ 를 얻는다.

둘째, 정규화된 모든 질의 영역에 대해서 각 축별 구간의 크기를 합산한 값의 비율로서 페이지 영역의 최적 구간비 $a:b:c:d$를 결정한다.

셋째, 최적 구간비에 최대한 가까운 페이지 영역으로 구성된 공간 액세스 구조를 구축한다. 여기서는 제 3.2절의 영역 분할 정책을 적용한다. 즉, 계속되는 공간 객체의 삽입으로 MBR-MLGF의 데이터 페이지에 오버플로우가 발생하면, 이 데이터 페이지에 대응하는 페이지 영역은 구간 이등분 정책을 사용하여 같은 크기의 두 영역으로 분할되고, 원 데이터 페이지의 객체들은 분할된 페이지 영역에 대응하는 두 개의 데이터 페이지로 나뉘어 저장된다. 이때 페이지 영역의 구간 이등분 정책으로 제 3.2절의 영역 분할 정책을 적용하는 것이다. 즉, 둘째 단계에서 결정된 최적 구간비($a:b:c:d$)와 같은 모양을 가지는 가상의 질의 영역($a\times b\times c\times d$)이 임의의 위치에 주어진다고 가정하고, 분할이 요구되는 페이지 영역($p(w)\times p(x)\times p(y)\times p(z)$)이 각 축에 대해 구간 이등분에 의한 분할 후의 한 페이지 영역과 교차하게 되는 질의 영역의 위치 범위의 크기(예를들어, W축의 구간을 이등분했을 때 그 크기는 $(p(w)/2+a)(p(x)+b)(p(y)+c)(p(z)+d)$이다.)를 각각 계

산한 다음 그 값이 가장 작게 되는 축을 분할 축으로 선택한다.

• 설계 정보

이차원 원공간에 주어진 질의패턴으로 사차원 변환공간(w, x, y, z 축으로 구성)상에 변환한 n개의 사차원 질의 영역

(1) 각 질의 영역의 형태 : $q_i(w) \times q_i(x) \times q_i(y) \times q_i(z)$ $(i=1, \cdots, n)$
(2) 각 질의 영역에 포함되는 공간 객체의 개수 : $n\_o_i (i=1, \cdots, n)$

• 알고리즘

단계 1 : 각 질의 영역의 정규화 $(i=1, \cdots, n)$

(1) 각 질의 영역의 객체 밀집도 d_i를 구한다.

$$d_i = \frac{n\_o_i}{q_i(w) \times q_i(x) \times q_i(y) \times q_i(z)}$$

(2) 밀집도 d_i로서 정규화된 질의 영역의 각 축의 크기를 구한다.

$$q_i'(w) = q_i(w) \times d_i^{1/4}$$
$$q_i'(x) = q_i(x) \times d_i^{1/4}$$
$$q_i'(y) = q_i(y) \times d_i^{1/4}$$
$$q_i'(z) = q_i(z) \times d_i^{1/4}$$

단계 2 : 페이지 영역의 최적 구간비 $(a:b:c:d)$ 결정

$$a:b:c:d = \sum_{i=1}^{n} q_i'(w) : \sum_{i=1}^{n} q_i'(x) : \sum_{i=1}^{n} q_i'(y) : \sum_{i=1}^{n} q_i'(z)$$

단계 3 : 최적 구간비에 가장 가까운 페이지 영역으로 구성된 공간 액세스 구조의 구축

(1) 공간 액세스 구조에 객체 삽입

(2) 데이터 페이지에 오버플로우가 발생하면, 데이터 페이지 분할
⇒ 대응하는 페이지 영역 $(p(w) \times p(x) \times p(y) \times p(z))$의 분할전략 : 다음 식들의 계산에 따른 결과로서 분할 축 결정

- $(p(w)/2+a)(p(x)+b)(p(y)+c)(p(z)+d)$의 값이 최소이면 W축 분할
- $(p(w)+a)(p(x)/2+b)(p(y)+c)(p(z)+d)$의 값이 최소이면 X축 분할
- $(p(w)+a)(p(x)+b)(p(y)/2+c)(p(z)+d)$의 값이 최소이면 Y축 분할
- $(p(w)+a)(p(x)+b)(p(y)+c)(p(z)/2+d)$의 값이 최소이면 Z축 분할

(3) 삽입할 객체가 있으면, (1), (2), (3) 반복

(그림 5) 이차원 원공간에 대한 공간 액세스 구조의 물리적 데이터베이스 설계 알고리즘

4. 성능 평가

본 절에서는 공간 액세스 구조의 물리적 데이터베이스 설계 알고리즘의 유용성을 다양한 실험을 통하여 제시한다. 실험의 목적은 공간 액세스 구조를 구성하는 페이지 영역의 모양에 대한 질의 영역의 모양과 크기, 그리고 저장된 객체들의 분포 등 여러 가지 인자들의 변화에 대하여 제안된 기법의 유용성을 실제 실험을 통하여 검증하는 것이다. 제 4.1절에서는 성능평가를 위하여 사용된 실험 환경에 대하여 기술하고, 제 4.2절에서는 실험 결과를 제시하고 이를 분석한다.

4.1 실험 환경

본 실험에서는 100,000개의 객체를 포함하는 두 종류의 MBR-MLGF를 구축하였다. 하나는 X축과 Y축으로 구성된 이차원 MBR-MLGF이고, 다른 하나는 W축, X축, Y축, 및 Z축으로 구성된 사차원 MBR-MLGF이다.

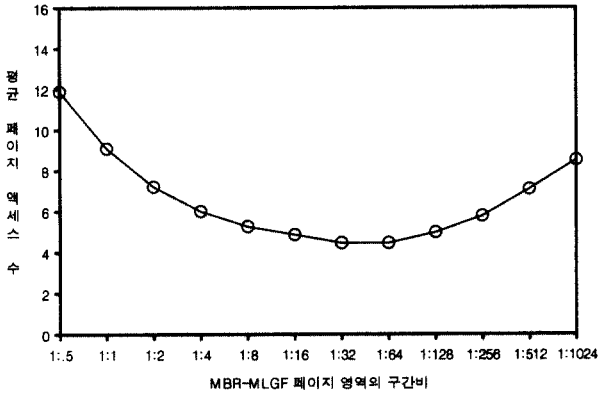
이차원 MBR-MLGF의 구축에 사용한 데이터의 분포 특성은 균일 분포와 비균일 분포로 구분한다. 균일 분포의 데이터는 각 축의 값이 $[-2^{31}, 2^{31}-1]$인 구간내에서 균일 분포하게 하고, 비균일 분포의 데이터로는 각 축의 값이 $[-2^{31}, 2^{31}-1]$인 구간내에서 표준 편차 σ가 $2^{31} \times 2/5$인 $N(0, \sigma^2)$의 정규 분포를 취하게 하여, 두 가지 분포 모두에 대하여 Y축의 값이 X축의 값보다 크거나 같은 경우만을 사용한다. 한 축의 값이 다른 축의 값보다 크거나 같은 경우만을 취한 이유는 본 논문에서 공간 액세스 구조로 구석점 변환기법을 사용하기 때문이다. 그리고, 사차원 MBR-MLGF의 구축에 사용한 데이터는 비균일 분포 데이터로서, 각 축의 값은 $[-2^{31}, 2^{31}-1]$의 구간내에서 표준 편차 σ가 $2^{31} \times 2/5$인 $N(0, \sigma^2)$의 정규 분포를 취하도록 하여, X축의 값은 W축의 값보다 크거나 같고 Z축의 값은 Y축의 값보다 크거나 같은 경우만을 사용한다.

질의 패턴의 구성을 위하여 사용한 질의 영역들의 형태는 이차원 질의 영역인 경우에는 질의 영역의 구간비가 1 : 1, 1 : 2, 1 : 4, 1 : 8, 1 : 16, 1 : 32, 1 : 64, 1 : 128, 및 1 : 256인 각각에 대해서, 질의 영역의 크기에 따라 다음과 같이 구성한다 : (1) 크기가 도메인 공간의 1/200로서 대영역(Large)인 L1, L2, L4, L8, L16, L32, L64, L128, 및 L256형태의 질의 영역, (2) 크기가 도메인 공간의 1/2000로서 중영역(Medium)인 M1, M2, M4, M8, M16, M32, M64, M128, 및 M256형태의 질의 영역, (3) 크기가 도메인 공간의 1/20000로서 소영역(Small)인 S1, S2, S4, S8, S16, S64, S128, 및 S256형태의 질의 영역 등이다. 그리고, 사차원 질의 영역인 경우에는 크기가 도메인 공간의 1/20000로서 소영역인 경우에만 한정하여 질의 영역의 구간비가 각각 1 : 1 : 1 : 1, 1 : 2 : 4 : 8, 1 : 4 : 16 : 64, 1 : 8 : 64 : 512, 및 1 : 16 : 256 : 4096인 S1_1_1_1, S1_2_4_8, S1_4_16_64, 및 S1_8_64_512, 및 1_16_256_4096 형태의 질의 영역 등을 사용한다.

4.2 실험 결과

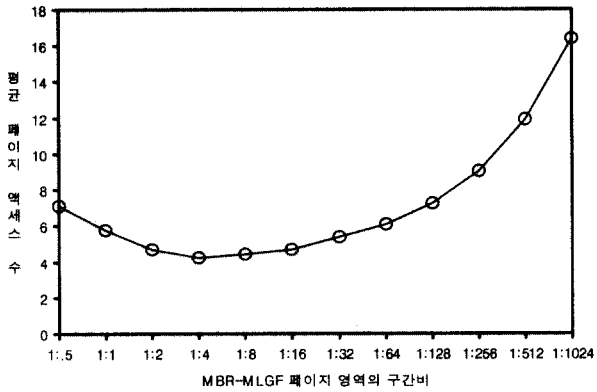
첫 번째 실험에서는 이차원 균일 분포의 데이터에 대하여 특정 구간비의 페이지 영역을 갖는 여러 개의 이차원 MBR-MLGF들을 생성하고, 각각에 대하여 다양한 질의 영역 형태를 갖는 여러 질의들로 구성된 질의 패턴을 처리할 때 발생하는 평균 페이지 접근 수를 측정한다. 실험에 사용된 질의 영역의 형태는 대영역인 L1, L2, L4, 중영역인 M8, M16, M32, 소영역인 S64, S128, S256 등이며, 질의 패턴을 구성하기 위하여 대영역 질의 형태는 10개씩 도메인 공간상의 중앙에 집중되도록 생성하고, 중영역 질의 형태는 100개씩 도메인 공

간상의 좌측하단에 집중되도록 생성하며, 대영역 질의 형태는 1000개씩 도메인 공간상의 우측상단에 집중되도록 생성한다. (그림 6)은 실험 결과를 그래프 형태로 나타낸 것이다. 모든 질의 영역들에 대하여 각 축의 구간 크기를 더한 값의 비로 계산된 페이지 영역의 최적 구간비는 1 : 56.7이며, (그림 6)에서 알 수 있는 바와 같이 이 비율과 가장 유사한 1 : 64를 페이지 영역의 구간비로 가지는 MBR-MLGF에서 가장 좋은 성능을 보인다. 이와 같은 실험결과는 데이터가 균일하게 분포하면, 주어진 다양한 형태의 질의 영역들에 의해 교차되는 페이지 영역의 개수는 공간 액세스 구조를 구성하는 페이지 영역의 구간비가 주어진 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산된 최적 구간비에 가까울수록 적어지기 때문이다.



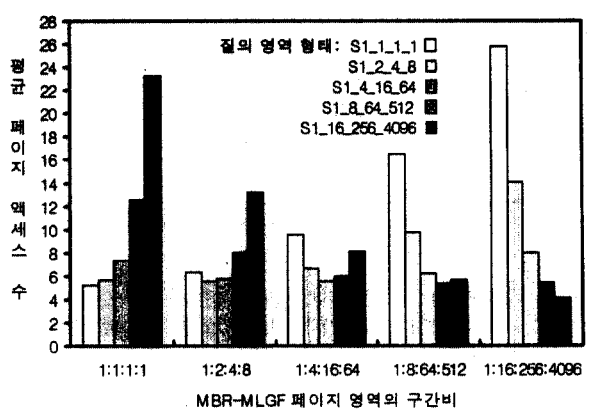
(그림 6) 이차원 균일 분포 데이터에 대한 특정 구간비의 페이지 영역을 갖는 MBR-MLGF별 혼합 질의 패턴의 질의처리 성능

두 번째 실험에서는 이차원 비균일 분포의 데이터에 대하여 첫 번째 실험과 동일한 실험을 수행한다. 즉, 이차원 비균일 분포의 데이터로서 특정 구간비의 페이지 영역을 갖는 여러 개의 이차원 MBR-MLGF들을 생성하고, 각각에 대하여 첫 번째 실험에 주어진 질의 패턴의 질의들을 처리할 때 발생하는 평균 페이지 접근 수를 측정한다. (그림 7)은 실험 결과를 그래프 형태로 나타낸 것이다. 정규화된 모든 질의 영역들에 대하여 각 축의 구간 크기를 더한 값의 비로 계산된 페이지 영역의 최적 구간비는 균일 분포의 데이터로 구성된 첫 번째 실험에서의 구간비 1 : 56.7과는 매우 다른 1 : 3.5로 계산되었으며, (그림 7)에서 알 수 있는 바와 같이 이 비율과 가장 유사한 1 : 4를 페이지 영역의 구간비로 가지는 MBR-MLGF에서 가장 좋은 성능을 보인다. 이와같은 실험 결과는 데이터가 비균일하게 분포하면, 주어진 다양한 형태의 질의 영역들에 의해 교차되는 페이지 영역의 개수는 페이지 영역의 구간비가 주어진 질의 영역들에 대해 정규화 과정을 통하여 각 축별로 구간 크기를 더한 값의 비로서 계산된 최적 구간비에 가까울수록 적어지기 때문이다.



(그림 7) 이차원 비균일 분포 데이터에 대한 특정 구간비의 페이지 영역을 갖는 MBR-MLGF별 혼합 질의 패턴의 질의처리 성능

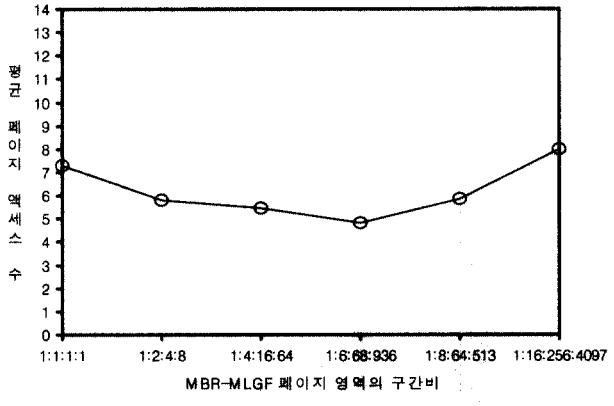
세 번째 실험에서는 사차원 비균일 분포를 가지는 데이터에 대하여 특정 구간비의 페이지 영역을 갖는 여러 개의 사차원 MBR-MLGF들을 생성하고, 각각에 대하여 고유의 질의 영역 형태를 갖는 여러 질의들로 구성된 질의패턴을 처리할 때 발생하는 페이지 접근 수를 측정한다. 이 실험의 목적은 사차원 비균일 분포 데이터에 대하여 도메인 공간상의 임의의 위치에 주어지는 특정 형태의 질의 영역에 의해 접근되는 평균 페이지의 개수는 MBR-MLGF의 페이지 영역의 구간비가 주어진 질의 영역의 구간비와 같게 될 때 최소로 됨을 확인하는 것이다. MBR-MLGF들의 구축에 사용된 특정 페이지 영역의 구간비는 1 : 1 : 1 : 1, 1 : 2 : 4 : 8, 1 : 4 : 16 : 64, 1 : 8 : 64 : 512, 및 1 : 16 : 256 : 4096 등 다섯 가지이며, 질의 영역의 형태는 S1_1_1_1, S1_2_4_8, S1_4_16_64, S1_8_64_512, 및 S1_16_256_4096 등의 다섯 가지이다. 이러한 각 질의 영역의 형태별로 1000개의 질의 영역을 도메인 공간상에 균일하게 생성하고, 이들 질의를 처리하는데 발생하는 평균 페이지 접근 수를 측정한다. (그림 8)은 이에 대한 실험 결과를 그래프로 나타낸 것이다. 모든 형태의 질의 영역에 대하여, 그



(그림 8) 사차원 비균일 분포의 데이터에 대한 특정 구간비의 페이지 영역을 갖는 MBR-MLGF에 대한 질의 영역의 형태별 질의 처리 성능

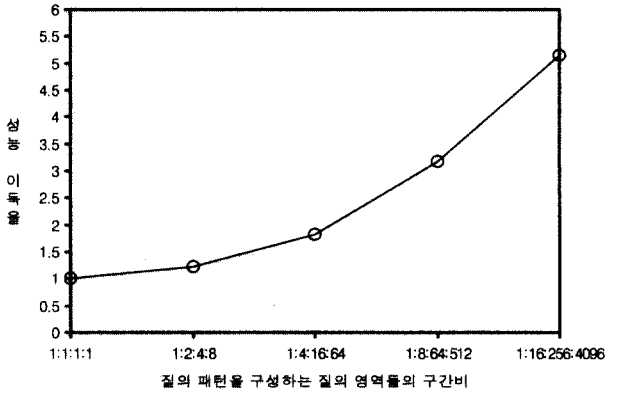
질의 영역의 구간비를 페이지 영역의 구간비로 가지는 MBR-MLGF에서 가장 좋은 성능을 보였다. 이와같은 실험 결과는 사차원 MBR-MLGF에 대해서도 본 본문에서 제안한 공간 액세스 구조의 물리적 데이터베이스 설계 기법이 유효함을 보이는 것이다.

네 번째 실험에서는 사차원 비균일 분포의 데이터에 대하여 두 번째 실험과 동일한 실험을 수행한다. 즉, 사차원 비균일 분포의 데이터로서 특정 구간비의 페이지 영역을 갖는 여러 개의 사차원 MBR-MLGF들을 생성하고, 각각에 대하여 다양한 질의 영역 형태를 갖는 여러 질의들로 구성된 질의 패턴을 처리할 때 발생하는 평균 페이지 접근 수를 측정한다. 실험에 사용된 질의 영역의 형태는 S1_1_1_1, S1_2_4_8, S1_4_16_64, S1_8_64_512, 및 S1_16_256_4096 등의 다섯 가지로, 질의 패턴을 구성하기 위하여 각각 200개씩 도메인 공간상에 균일하게 분포하도록 한다. (그림 9)는 실험 결과를 그래프 형태로 나타낸 것이다. 정규화된 모든 질의 영역들에 대하여 각 축의 구간 크기를 더한 값의 비는 1 : 6 : 68 : 936으로 계산되었으며, (그림 9)에서 알 수 있는 바와 같이 이 비율과 같은 구간비의 페이지 영역을 갖는 MBR-MLGF에서 가장 좋은 성능을 보인다. 이와같은 실험결과는 사차원의 MBR-MLGF에 대해서도 두 번째 실험의 이차원 MBR-MLGF에서와 마찬가지로, 다양한 형태의 질의 영역들에 의해 교차되는 페이지 영역의 개수를 최소로 하는 페이지 영역의 최적 구간비는 정규화 과정을 통하여 주어진 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 계산할 수 있음을 보이기 위한 것이다. 또한, 이 실험의 결과는 제 3.3절에서 제시한 이차원 원공간에 대하여 변환공간으로 일반화한 공간 액세스 구조의 물리적 데이터베이스 설계 알고리즘의 실용성을 입증하는 것이다.



(그림 9) 사차원 비균일 분포 데이터에 대한 특정 구간비의 페이지 영역을 갖는 MBR-MLGF별 혼합 질의 패턴의 질의처리 성능

끝으로, 다섯 번째 실험에서는 이차원 원공간에 대하여 변환기법에 의한 공간 액세스 구조의 물리적 데이터베이스 설계 기법을 이용하여 구성한 MBR-MLGF가 기존의 순환 분할전략(즉, 구간비 = 1 : 1 : 1 : 1)으로 구성한 MBR-MLGF 사이의 성능을 비교한다. 먼저, 다섯 가지의 사차원 질의 영역의 형태인 S1_1_1_1, S1_2_4_8, S1_4_16_64, S1_8_64_512, 및 S1_16_256_4096에 대하여, 각 형태별로 1000개의 질의 영역들이 도메인 공간상에 균일하게 주어지는 다섯 가지의 질의 패턴을 생성한다. 그리고, 각 질의 패턴에 대하여 최적의 구간비(질의 패턴을 구성하는 질의 영역들의 구간비와 동일)를 갖는 페이지 영역들로 구성된 MBR-MLGF를 생성하여 그 질의 패턴을 처리할 때 발생하는 평균 페이지 접근 수를 구하고, 이 값에 대한 구간비가 1 : 1 : 1 : 1인 페이지 영역들로 구성된 MBR-MLGF에서 같은 질의 패턴을 처리할 때 발생하는 평균 페이지 접근 수의 비율을 측정한다. (그림 10)은 이에 대한 실험 결과를 나타낸 것이다. 가로축은 각 질의 패턴을 구성하는 질의 영역들의 구간비를 나타내며, 세로축은 제안된 기법을 사용하는 경우의 성능 이득이 몇 배인가를 나타낸다.



(그림 10) 질의 패턴을 구성하는 질의 영역의 구간비별 공간 액세스 구조의 물리적 데이터베이스 설계기법에 의해서 생성된 사차원 MBR-MLGF의 성능 효율

(그림 10)에서 나타난 바와 같이 질의 영역의 구간비가 1 : 1 : 1 : 1에서 멀어질수록 제안된 기법을 사용하는 경우의 성능 개선 효과가 뚜렷해짐을 볼 수 있다. 즉, 질의 영역의 구간비가 1 : 16 : 256 : 4096인 경우 질의처리 성능이 5.14배까지 향상됨을 볼 수 있으며, 구간비가 더 커질수록 더욱더 향상될 수 있음을 나타낸다. 이러한 결과는 제 3.3절에서 제시한 사차원공간 액세스 구조의 물리적 데이터베이스 설계 기법의 성능 개선 효과를 잘 나타내는 것이다.

5. 결 론

본 논문에서는 변환기법을 이용한 공간 액세스 구조에 대하여 주어진 공간 질의의 패턴에 관한 정보로서 질의 처리의 성능을 최적으로 보장할 수 있는 공간 액세스 구조의 물리적 데이터베이스 설계기법을 제시하였다. 변환기법을 이용한 공간 액세스 구조에서는 모든 객체들을 변환공간상의 점 객체

들로 표현하기 위하여 변환공간의 차원을 원공간 차원의 두 배로 한다. 먼저, 원공간의 다양한 공간 질의들은 변환공간에서는 한 가지 형태의 범위 질의들로 변환하여 처리할 수 있음을 보이고, 이러한 범위 질의들을 처리하기 위하여 접근하는 페이지의 개수를 최소로 하는 최적의 공간 액세스 구조를 구축한다.

본 논문에서 제안한 공간 액세스 구조의 물리적 데이터베이스 설계 기법은 공간 액세스 구조를 구성하는 변환공간의 분할 상태를 나타내는 페이지 영역들의 구간비가 변환공간상의 범위 질의가 위치하는 질의 영역의 구간비와 일치할 때, 질의 처리 시에 발생되는 페이지 접근 수가 최소로 되는 성질을 바탕으로 한다. 제안한 설계기법에서는 먼저, 도메인 공간상의 데이터 분포 특성을 고려하기 위하여 질의 패턴에 주어진 각 질의 영역에 대하여 정규화 과정을 거친다. 그리고, 정규화된 모든 질의 영역을 대상으로 각 축별 구간의 크기를 합산한 값의 비율로서 페이지 영역의 최적 구간비를 결정하고, 이 최적 구간비에 최대한 가까운 모양의 페이지 영역 구성된 공간 액세스 구조를 구축한다.

또한, 본 논문에서는 이와 같은 공간 액세스 구조의 물리적 데이터베이스 설계기법의 성능평가를 위하여, 변환기법을 이용하는 공간 액세스 구조의 하나인 MBR-MLGF를 대상으로 페이지 영역의 모양이 최적 구간비에 근접하도록 하는 영역 분할전략을 제시하고, 이를 이용하여 다양한 실험을 수행하였다. 실험 결과에 의하면, 주어진 질의 패턴과 데이터 분포에 따라 최적의 MBR-MLGF를 구성할 수 있었으며, 이차원 원공간에 대한 사차원 변환공간상의 질의 영역의 모양이 편향된 정도에 따라 기존의 정방형(구간비가 1 : 1 : 1 : 1인 경우) 모양의 페이지 영역으로 구성된 MBR-MLGF에 비해 질의처리의 성능이 (그림 10)에서와 같이 급격히 향상되는 것으로 나타났다. 특히, 질의 영역의 구간비가 1 : 16 : 256 : 4096인 경우에는 질의처리 성능이 5.14배까지 향상됨을 볼 수 있었다. 이것은 제안된 기법이 실제적으로 매우 유용함을 보여주는 것이다.

참 고 문 헌

[1] Guting, O., "An Introduction to Spatial Database Systems," *The VLDB Journal*, Vol.3, No.4, pp.357-399, Oct., 1994.

[2] Song, J. W., Whang, K. Y., and Kim, S. W., "Spatial Join Processing Using Corner Transformation," *IEEE Trans. on Knowledge and Data Engineering*, Vol.11, No.4, Aug., 1999.

[3] Orenstein, J., "Spatial Query Processing in an Object-Oriented Database System," In *Proc. int'l Conf. on Management of Data*, ACM SIGMOD, pp.326-336, 1986.

[4] Faloutsos, C., "Gray Codes for Partial Match and Range Queries," *IEEE Trans. on Software Engineering*, Vol.14, No.10, pp.1381-1393, Oct., 1988.

[5] Samet, H., *Applications of Spatial Data Structures : Computer Graphics, Image Processing and GIS*, Addision-Wesley, 1990.

[6] Sellis, T. et al., "The R^+-tree : A Dynamic Index for Multidimensional Objects," In *Proc. 13th int'l Conf. on Very Large Data Bases*, pp.507-518, 1987.

[7] Guttman, K., "R-trees : A Dynamic Index Structure for Spatial Searching," In *Proc. int'l Conf. on Management of Data*, ACM SIGMOD, pp.47-57, 1984.

[8] Beckmann, N., Kriegel, H. P., and Schneider, R., "The R^*-tree : An Efficient and Robust Access Method for Points and Rectangles," In *Proc. int'l Conf. on Management of Data*, ACM SIGMOD, pp.322-331, 1990.

[9] Ooi, B. C. et al., "Spatial Indexing in Binary Decomposition and Spatial Bounding," *Information Systems*, Vol.16, No.2, pp.211-237, 1991.

[10] Hinrichs, K. and Nievergelt, J., "The Grid File : A Data Structure Designed to Support Proximity Queries on Spatial Objects," In *Proc. int'l Workshop on Graph Theoretic Concepts in Computer Science*, pp.100-113, 1983.

[11] Seeger, B. and Kriegal, H. P., "Techniques for Design and Implementation of Efficient Spatial Access Methods," In *Proc. 14th int'l Conf. on Very Large Data Bases*, pp.360-371, 1988.

[12] Pagel, B. U. et al., "The Transformation Technique for Spatial Objects Revisited," In *Proc. 3rd int'l Symp. on Spatial Databases(SSD'93)*, 1993.

[13] Lu, H. and Ooi, B. C., "Spatial Indexing : Past and Future," *IEEE Data Engineering Bulletin*, Vol.16, No.3, pp.16-21, Sept., 1993.

[14] Chang, J. M. and Fu, K. S., "A Dynamical Clustering Technique for Physical Database Design," In *Proc. int'l Conf. on Management of Data*, ACM SIGMOD, pp.183-199, Santa Monica, May, 1980.

[15] Whang, K. Y. et al., "Separability-An Approach to Physical Database Design," *IEEE Trans. on Computers*, Vol.c-33, No.3, pp.209-222, Mar., 1984.

[16] Finkelstein, S. et al., "Physical Database Design for Relational Databases," *ACM Trans. on Database Systems*, Vol. 13, No.1, pp.91-128, Mar., 1988.

[17] Elmasri, R. and Navathe, S. B., *Fundamentals of Database Systems*, Benjamin/ Cummings Publishing Co., Redwood City, California, Second Ed., 1994.

[18] Yu, C. T. et al., "Adaptive Record Clustering," *ACM Trans. on Database Systems*, Vol.10, No.2, pp.180-204, June, 1985.

[19] Whang, K. Y. and Krishnamurthy, R., *Multilevel Grid File*, IBM Research Report RC 11516, 1985.

[20] Whang, K. Y., Kim, S. W., and Wiederhold, G., "Dynamic Maintenance of Data Distribution for Selectivity Estimation," *The VLDB Journal*, Vol.3, No.1, pp.29-51, Jan., 1994.

[21] Kriegel, H. P., "Query Processing in Spatial Database Systems," In Book *New Results and Trends in Computer Science*, Lecture Notes in Computer Science 555, Springer Verlag, pp.172-191, 1991.

[22] Lee, J. H. et al., "A Physical Database Design Method for Multidimensional File Organizations," Information Sciences, Vol.102, No.3, pp.31-65, 1997.

이 종 학

e-mail : jhlee11@cuth.cataegu.ac.kr

1982년 경북대학교 전자공학과(전자계산 전공) 졸업(학사)

1984년 한국과학기술원 전산학과 졸업(공학석사)

1997년 한국과학기술원 전산학과 졸업(공학박사)

1991년 정보처리기술사

1984년~1987년 금성통신(주) 부설연구소 주임연구원

1987년~1998년 한국통신 연구개발본부 선임연구원

1998년~현재 대구효성가톨릭대학교 컴퓨터정보통신공학부 조교수

관심분야 : 데이타베이스 시스템, 객체지향 데이타베이스, 트랜잭션 프로세싱, 지리정보 시스템 등

박 병 권

e-mail : bpark@daunet.donga.ac.kr

1986년 서울대학교 산업공학과 학사

1988년 한국과학기술원 경영과학과 석사

1998년 한국과학기술원 전산학과 박사

1988년~1993년 삼성전자 컴퓨터개발실 주임연구원

1998년~2000년 삼성전자 중앙연구소 소프트웨어센터 선임연구원

2000년~현재 동아대학교 경영정보과학부 전임강사

관심분야 : 데이터베이스, 데이터웨어하우스, 정보검색, 전자상거래