질의패턴에 따른 다차원 파일구조의 구성방법

이정아, 이종학
대구가톨릭대학교 컴퓨터정보통신공학부

e-mail:{leejunga, jhlee11}@cu.ac.kr

Organization of Multidimensional File Structures Defending on a Query Pattern

Jung-A Lee, Jong-Hak Lee
School of Computer and Information Communications Engineering, Catholic Univ. of Daegu

요 약

본 논문에서는 다차원 파일구조를 주어진 질의 패턴에 의해 최적으로 구성할 수 있는 방법을 제시한다. 지금까지의 다차원 파일구조는 응용 시스템에서 주어지는 질의의 패턴을 고려하지 않고 다차원 파일구조를 구성하는 애트리뷰트들의 클러스터링 정도를 동일하게 취급하였다. 그러나 다차원 파일구조를 이용하는 대부분의 응용 시스템에서 구성 애트리뷰트들 사이의 액세스 정도를 크게 다르게 하는 질의 패턴을 보인다. 따라서 본 논문에서는 다차원 파일구조의 응용 시스템에서 주어지는 질의 정보를 이용하여 각 구성 애트리뷰트들 사이의 클러스터링 정도를 각각 다르게 반영함으로써 최적이 되는 다차원 파일구조를 구성하는 방안을 제시한다. 먼저 질의처리의 성능이 질의 패턴에 주어진 질의 영역의 모양과 다차원 파일구조의 도메인 공간의 분할 상태를 나타내는 페이지 영역의 모양 사이의 유사성에 따라 크게 영향 받음을 보이고, 이러한 특성을 이용하여 수학적 분석을 통하여 제안된 기법의 이론적인 배경을 증명한다.

1. 서론

다차원 파일구조는 다차원 클러스터링(multidimensional clustering)[3]을 지원하는 파일구조로서, 여러 개의 애트리뷰트로 구성된 질의를 효과적으로 처리할 수 있다. 이미 다차원 파일구조에 대한 많은 연구가 진행되어 왔으며, 대표적인 예로는 KD-트리, K-D-B-트리[7, 8, 10], 4D-트리, LSD-트리, hB-트리[1] 및 MD-트리[5] 등을 비롯하여 다차원 선형 해싱(multidimensional linear hashing), 다차원 신장 해싱(multidimensional extendible hashing), 그리드 파일(grid file), BANG-파일, 계층 그리드 파일(multilevel grid file)[9]등이 있다.

효과적인 클러스터링을 위해서는 레코드들을 그룹화하여 페이지 단위로 저장할 때, 질의처리 시에 액세스되는 전체 페이지의 개수를 최소화하는 방안을 고려하여야 한다. 즉, 빈번히 함께 액세스되는 레코드들을 같은 페이지 내에 저장함으로써 질의처리 시 액세스되는 페이지의 개수를 최소화 하는 것이 필요하다[2, 4, 11].

여러 애트리뷰트들이 클러스터링 특성을 공유하는 다차원 파일구조에서는 영역 분할전략(region splitting strategy)의 변화를 통하여 애트리뷰트 별로 클러스터링의 정도를 조정할 수 있다. 기존의 다차원 파일구조에서는 영역의 분할 시 주로 순환 분할전략(cyclic spitting strategy)[6, 9]을 사용하도록 하여 각 페이지와 대응되는 영역의 형태가 정방형이 되도록 함으로써 모든 애트리뷰트들이 클러스터링 특성을 같은 정도로 공유하도록 하고 있다.

그러나, 사용자가 요구하는 질의에서는 일반적으로 다차원 파일구조를 구성하는 각 애트리뷰트에 따라 질의 조건의 구간 크기가 다르며, 큰 구간의 질의 조건이 특정 애트리뷰트에 편향되게 주어지는 경향이 있다. 따라서, 이러한 경우 모든 애트리뷰트들이 클러스터링 특성을 같은 정도로 공유하게 되면 질의처리시의 성능이 저하된다.

본 논문에서는 미리 주어진 사용자 질의 정보를 기반으로 질의처리 비용을 최소화하는 다차원 파일구조의 구성방법을 제시한다. 제안된 기법은 사용자 질의 패턴을 사전에 분석함으로써 질의 처리시 발생하는 페이지 액세스 수를 최소화할 수 있는 최적의 애트리뷰트별 클러스터링 정도를 구하는 것이다. 수학적 분석을 통하여 제안된 기법의 이론적인 배경을 증명한다. 본 논문에서 제안하는 다차원 파일구조의 구성방법은 다차원 파일구조를 이용하는 많은 응용에서 성능을 개선하는 좋은 해결책을 제시한다는 면에서 큰 의미를 갖는다.

2. 다차원 파일구조의 특징

파일은 속성들의 리스트로 구성된 레코드들의 모임이다. 파일의 구조는 레코드를 구성하는 속성들 중에서 일부에 의해서 결정되며, 이와 같이 파일을 구성하는데 참여하는 속성들을 구성 속성(organizing attribute)이라 한다. 그리고, 두 개 이상의 구성 속성을 가지는 파일구조를 다차원 파일구조라 한다. 도메인은 한 속성이 취할 수 있는 모든 값들의 집합

이며, 다차원 파일구조 내에서 하나의 축에 해당된다. 모든 속성에 대한 도메인들의 카티젼 곱(Cartesian product)을 도메인 공간(domain space)이라 정의하고, 도메인공간의 일부분을 영역(region)이라 한다.

다차원 파일구조는 영역의 분할시 분할 경계를 정하는 기준에 따라 크게 두 가지로 분류된다. 하나는 영역에 포함된 데이터의 개수를 균등하게 양분하는 속성 값을 분할 경계로 하는 데이터 기준 분할전략의 파일구조[5, 8]로서, 분할 경계가 데이터의 분포에 따라 유동적이다. 또 다른 하나는 영역의 크기를 반분하는 위치를 분할 경계로 하는 영역 기준 분할전략의 파일구조[6, 9]로서, 분할경계가 데이터의 분포에 상관없이 고정적이다. 영역 기준 분할전략을 사용하는 파일구조는 영역 분할의 경계가 항상 고정되므로, 파일구조가 데이터의 입력 순서에 관계없이 일정하게 유지되고, 영역 분할전략에 따라 영역의 모양을 쉽게 조정할 수 있는 특징이 있다. 본 논문에서 제안하는 다차원 파일구조의 구성방법은 이러한 영역 기준분할전략을 사용하는 기법들의 특징을 기반으로 한다.

3. 문제 정의

본 논문에서는 일련의 사용자 질의 패턴에 대해서 평균 질의처리 비용을 최소화할 수 있는 다차원 파일구조를 구성하기 위하여, 다차원 파일구조의 도메인 공간을 구성하는 페이지 영역들의 최적의 구간비를 결정하는 문제를 다차원 파일구조의 최적 구성방법이라 한다. 지금까지의 다차원 파일구조의 구성에서는 각 축을 미리 결정된 일정한 순서에 의해서 번갈아 가면서 분할한다. 이러한 방식은 사용자로부터 주어지는 질의 패턴의 특성을 전혀 반영하지 않은 것이다.

다차원 파일구조에 대한 사용자 질의는 검색 결과에 포함되는 레코드들이 만족해야할 술어(Predicate)들의 결합(conjunction)으로 구성된다. 완전부합 질의(exact match query)는 모든 속성에 대해서 등식 술어(즉, 속성 A = 'a')들의 결합으로 정의한다. 그리고 부분부합 질의(partial match query)는 속성들 중에서 일부만이 술어에 포함된 질의이며, 범위 질의(range query)는 등식 술어들과 함께 적어도 하나 이상의 속성에 의한 범위 술어(즉, 속성 A > 'a')들이 결합된 질의이다[9].

본 논문에서는 Robinson[8]에 의해서 정의된 다음과 같은 질의 영역이란 용어를 사용한다. "사용자가 요구하는 모든 질의는 도메인 공간내의 영역들로 표현할 수 있다. 이 영역은 파일을 구성하는 속성에 대한 구간들의 곱으로 표현되며, 이를 질의 영역이라 한다." 이 정의에 의하면, 완전부합 질의는 모든 속성들에 대한 구간이 단일 값으로 주어지는 점의 질의 영역으로 표현된 경우이고, 부분부합 질의는 일부 속성들에 대한 구간이 전체 도메인으로 주어지는 질의 영역으로 표현된 경우이다. 그리고 범위 질의는 적어도 하나 이상의 속성에 대한 구간이 전체 도메인의 일부 구간으로 주어지는 질의 영역으로 표현된 경우이다.

따라서, 모든 사용자 질의를 도메인 공간내의 질의 영역에 포함되는 레코드들을 탐색하는 연산으로 해석할 수 있다. 그러므로 다차원 파일구조의 구성방법의 문제는 일련의 사용자 질의 영역들에 의해서 교차되는 페이지 영역들의 개수를 최소화 하는 문제가 된다. 사용자 질의 패턴에 나타나는 질의 영역들의 형태에 대한 정보를 기반으로 질의 영역들에 의해 교차하는 페이지 영역들의 개수가 최소로 되는 최적 페이지 영역의 구간비를 결정하고, 가능한 이와 같은 구간비를 갖는 페이지 영역들이 되도록 하는 영역 분할 전략을 사용함으로써 최적의 다차원 파일구조를 구성할 수 있다.

4. 구성 원리

본 절에서는 설명의 편의를 위하여 속성이 두 개인 이차원 파일구조에 대해서, 질의 영역과 페이지 영역의 모양간의 상호관계에 의한 질의처리의 최적 조건으로서 다차원 파일구조의 구성 원리를 제시한다. 다양한 형태의 여러 질의 영역들로 주어진 질의 패턴에 대해서, 먼저 데이터가 균일하게 분포한다는 조건하에 질의처리의 최적 조건이 되는 최적 페이지 영역의 구간비를 결정할 수 있는 방법을 제시한다. 그리고 데이터가 비균일하게 분포하는 일반적인 경우로 확장한다.

4.1 균일 분포에서의 최적 페이지 영역의 구간비

데이터가 균일하게 분포하면, 도메인 공간의 위치에 상관없이 데이터의 밀집도(density)가 일정함으로 인하여 모든 페이지 영역의 크기가 일정하게 된다. 아래 [정리 1]은 도메인 공간을 구성하는 페이지 영역들의 크기가 일정할 때, 질의 패턴을 이루는 다양한 형태의 질의 영역들에 의해 교차되는 페이지 영역들의 총 개수를 최소로 하는 최적 페이지 영역의 구간비는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 결정할 수 있음을 정리한 것이다.

[보조정리 1] 크기가 $p(x)$로 일정한 페이지 구간(일차원 페이지 영역)으로 나누어져 있는 일차원 도메인 공간상에서, 크기가 $q(x)$인 임의의 질의 구간과 교차하게 되는 페이지 구간의 평균 개수는 $(\frac{q(x)}{p(x)+1})$개이다.

증명: 질의 구간의 크기 $q(x)$를 $np(x)+a$(n은 자연수, $0 \le a < p(x)$)라 하면, 질의 구간과 교차하게 되는 페이지 구간의 개수는 질의 구간이 시작되는 위치가 페이지 구간의 처음부터 $p(x)-a$까지의 범위에 오면 n+1개이고, 나머지 범위에 오면 n+2개이다. 따라서, 크기가 $q(x)$인 질의 구간과 교차하게 되는 페이지 구간의 평균 개수는

$$\frac{p(x)-a}{p(x)}(n+1)+\frac{p(x)-(p(x)-a)}{p(x)}(n+2)$$
$$= \frac{np(x)+a}{p(x)}+1 = \frac{q(x)}{p(x)}+1 \text{ 이다.} \quad \square$$

[보조정리 2] 크기가 $p(x)\times p(y)$로 일정한 페이지 영역들로 나누어져 있는 이차원 도메인 공간상에서 임의의 질의 영역 $q(x)\times q(y)$와 교차하게 되는 페이지 영역의 평균 개수는 $(\frac{q(x)}{p(x)+1})(\frac{q(y)}{p(y)+1})$개가 된다.

증명: 보조정리 2에 의하여 한축에 대해 교차하게 되는 페이지 영역의 평균 개수는$\frac{q(x)}{p(x)+1}$개이고, 또 다른 축에 대해 교차하게 되는 페이지 영역의 평균 개수는 $\frac{q(y)}{p(y)+1}$개이다. 따라서 그리드 형태의 페이지 영역들로 구성된 이차원 도메인 공간상에서, 이차원 질의 영역에 의해서 교차되는 페이지 영역의 평균 개수는 두 축간에 서로 독립적이므로$(\frac{q(x)}{p(x)+1})(\frac{q(y)}{p(y)+1})$개가 된다. $\square$

[정리 1] 크기가 $p(x)\times p(y)$로 일정한 페이지 영역들로 나누어져 있는 이차원 도메인 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x)\times q_i(y)(i=1,...,n)$에 대해 각 질의 영역과 교차하게 되는 페이지 영역의 총 개수를 최소로 하는 최적 페이지 영역의 구간비 $p(x):p(y)=$

$\sum_{i=1}^{n} q_i(x) : \sum_{i=1}^{n} q_i(y)$이다.

증명: 이차원 도메인 공간을 이루는 페이지 영역의 크기를

$$p(x) \times p(y) = B \quad (1)$$

라 할 때, [보조정리 1]에 의해 n개의 질의 영역 $p_i(x) \times p_i(y)(i = 1,...,n)$각각의 질의영역과 교차하게 되는 페이지 영역의 총 개수 $NB(p(x), p(y))$는 다음 식과 같다.

$$NB(p(x),p(y)) = \sum_{i=1}^{n} (\frac{q_i(x)}{p(x)} + 1)(\frac{q_i(y)}{p(y)} + 1) \quad (2)$$

수식(1)로부터 $p(y) = \frac{B}{p(x)}$이므로,

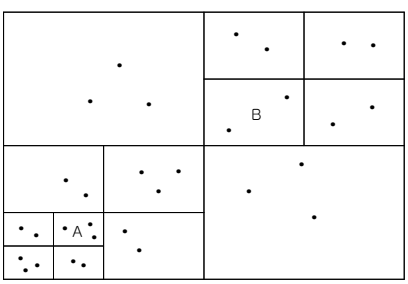
$$NB(p(x), \frac{B}{p(x)}) = \sum_{i=1}^{n} (\frac{q_i(x)}{p(x)} + 1)(\frac{q_i(y)p(x)}{B} + 1)$$

$$= \frac{\sum_{i=1}^{n} q_i(x)q_i(y)}{B} + \frac{\sum_{i=1}^{n} q_i(x)}{p(x)} + \frac{\sum_{i=1}^{n} q_i(y)p(x)}{B} + n \quad (3)$$

따라서, 수식 (3)의 값을 최소로 하는 $p(x)$를 구하면, $p(x) = \sqrt{(\sum_{i=1}^{n} q_i(x) / \sum_{i=1}^{n} q_i(y))B}$ 이고, 이러한 $p(x)$에 대한 $p(y)$는 수식 (1)에 의하여 $p(y) = \sqrt{(\sum_{i=1}^{n} q_i(y) / \sum_{i=1}^{n} q_i(x))B}$이다. 그러므로, $NB(p(x), p(y))$를 최소로 하는 최적 페이지 영역의 구간비 $p(x) : p(y) = \sum_{i=1}^{n} q_i(x) : \sum_{i=1}^{n} q_i(y)$이다. □

4.2 비균일 분포에서의 최적 페이지 영역의 구간비

도메인 공간내에서 데이터가 비균일하게 분포한다는 것은 도메인 공간내의 위치에 따라 레코드의 밀집도가 다름을 의미한다. 따라서, 이러한 경우에는 도메인 공간의 위치에 따라 페이지 영역의 크기가 달라진다. 즉, 레코드의 밀집도가 높은 곳에서는 밀집도가 낮은 곳에 비하여 많은 페이지가 할당되므로 각 페이지 영역의 크기는 작아지게 된다. 예를 들어 (그림 1)은 레코드의 밀집도가 상대적으로 높은 곳에 있는 페이지 영역 A의 크기가 레코드의 밀집도가 상대적으로 낮은 곳에 있는 페이지 영역 B의 크기보다 작은 것을 보여준다. 실제 데이터베이스 환경에서는 레코드가 도메인 공간내에서 균일하게 분포되지 않으므로, 본 절에서는 [정리 1]을 확장하여 레코드가 비균일하게 분포하는 경우에 대해서 다양한 형태의 질의 영역들에 의해 교차하는 페이지 영역들의 개수가 최소로 되는 최적 페이지 영역의 구간비를 결정할 수 있는 방법을 제시한다.



(그림 1) 비균일 데이터 분포에 따른 페이지 영역들의 크기 비교.

먼저, 레코드들이 균일하게 분포하는 경우의 [정리 1]에 대해 살펴보자. [정리 1]에서는 질의패턴을 구성하는 모든 질의 영역들에 대해 각 축별로 구간 크기를 더한 값의 비로서 결정한다. 예를 들어, 20×100인 형태 A의 질의 영역과 5 × 2인 형태 B의 질의 영역이 같은 비율로 주어질 때 최적 페이지 영역의 구간비는 25:102(즉, 20+5 : 100+2)로 결정된다. 여기에서 최적 페이지 영역의 구간비는 B형태 질의 영역의 구간비인 5:2보다 A형태 질의 영역의 구간비인 20:100에 더 가깝게 됨을 알 수 있다. 그 이유는 A형태의 질의 영역이 B형태의 질의 영역보다 크므로 질의처리 비용도 커지며, 이 결과 최적 페이지 영역의 구간비를 결정하는데 더 큰 영향을 미치기 때문이다. 즉, 균일분포 하에서는 각 질의 영역에 의해 교차되는 페이지 영역의 개수가 그 질의 영역의 크기에 비례하기 때문이다.

그러나, 데이터가 도메인 공간상에서 비균일하게 분포하는 경우에는, 질의 영역에 의해 교차되는 페이지 영역의 개수는 질의 영역의 크기뿐만 아니라 질의 영역이 주어진 위치의 데이터 밀집도에도 비례하게 되므로, [정리 1]에서와 같이 최적 페이지 영역의 구간비를 모든 질의 영역의 각 축별로 구간 크기를 단순히 더한 값의 비로서 구할 수 없다. 따라서, 이와 같은 경우에는 각 질의 영역의 크기에 대해 위치에 따른 데이터 밀집도를 가중치(weight)로 곱하여야 한다. 이와 같이 각 질의 영역의 구간비는 원래대로 유지하면서 크기에 대해 데이터 밀집도를 가중치로 곱한 질의 영역의 형태를 정규화된 질의 영역이라 하고, 정규화된 질의 영역으로 변환하는 과정을 질의 영역의 정규화(normalization)라 한다.

n개의 레코드를 포함하는 질의 영역 $q(x) \times q(y)$는 데이터 밀집도 d를 $\frac{n}{q(x) \times q(y)}$으로 하여 각 속성의 구간에 가중치 $\sqrt{d}$를 곱함으로써 정규화된 질의 영역 $q(x)\sqrt{d} \times q(y)\sqrt{d}$로 정규화 할 수 있다. 즉, 정규화된 질의 영역의 구간비는 $q(x)\sqrt{d} : q(y)\sqrt{d} = q(x) : q(y)$로서 원래 질의 영역의 구간비와 같게 유지되며, 크기만 원래 크기에 데이터 밀집도 d를 곱한 $q(x) \times q(y) \times d$가 된다. 여기에서, 정규화된 질의 영역의 크기는 그 질의 영역에 포함된 레코드의 개수와 같게 됨을 알 수 있다. 아래 [정리 2]는 데이터가 도메인 공간내에서 비균일하게 분포하는 경우에 최적 페이지 영역의 구간비를 정규화된 질의 영역들의 각 축별로 구간 크기를 더한 값의 비로서 결정할 수 있음을 입증한다.

[정리 2] 서로 다른 크기의 페이지 영역들로 나누어져 있는 이차원 도메인 공간상에서, 임의의 위치에 주어지는 n개의 질의 영역 $q_i(x) \times q_i(y)(i = 1,...,n)$에 대해 각 질의 영역의 데이터 밀집도를 d_i라 할 때, 각 질의 영역과 교차하게 되는 페이지 영역의 총 개수를 최소로 하는 최적 페이지 영역의 구간비 $p(x) : p(y) = \sum_{i=1}^{n} q_i(x)\sqrt{d_i} : \sum_{i=1}^{n} q_i(y)\sqrt{d_i}$이다.

증명: 이차원 도메인 공간을 이루는 서로 다른 크기의 모든 페이지 영역에 대해서, 정규화된 페이지 영역의 크기는 페이지 영역에 포함된 레코드의 개수(K)로서 모두 일정하다. 따라서 크기와 구간비가 일정한 정규화된 각 페이지 영역을 $p'(x) \times p'(y)$ 라 하면,

$$p'(x) \times p'(y) = K \quad (4)$$

가 된다. 이와 같이 정규화된 페이지 영역들은 크기와 구간비가 같으므로 [보조정리 2]를 이용하여, n개의 정규화된 질의 영역 $q_i(x)\sqrt{d_i} \times q_i(y)\sqrt{d_i}(i = 1, \cdots, n)$ 각각의 질의 영역과 교차하게 되는 정규화된 페이지 영역의 총 개수 NB(p'(x), p'(y))는 다음 식과 같이 유도된다.

$$NB(p'(x),p'(y))=\sum_{i=1}^{n}(\frac{q_i(x)\sqrt{d_i}}{p'(x)}+1)(\frac{q_i(y)\sqrt{d_i}}{p'(y)}+1) \quad (5)$$

수식 (4)로부터 $p'(y)=\frac{K}{p'(x)}$이므로,

$$NB(p'(x),\frac{K}{p'(x)})$$
$$=\sum_{i=1}^{n}(\frac{q_i(x)\sqrt{d_i}}{p'(x)}+1)(\frac{q_i(y)\sqrt{d_i}p'(x)}{K}+1)$$
$$=\frac{\sum_{i=1}^{n}q_i(x)q_i(y)d_i}{K}+\frac{\sum_{i=1}^{n}q_i(x)\sqrt{d_i}}{p'(x)}+\frac{\sum_{i=1}^{n}q_i(y)\sqrt{d_i}p'(x)}{K}+n \quad (6)$$

따라서, 수식 (6)의 값을 최소로 하는 $p'(x)$를 구하면, $p'(y)=\sqrt{(\sum_{i=1}^{n}q_i(x)\sqrt{d_i}/\sum_{i=1}^{n}q_i(y)\sqrt{d_i}K}$이고, 이러한 $p'(x)$에 대한 $p'(y)$는 수식 (4)에 의하여 $p'(y)=\sqrt{(\sum_{i=1}^{n}q_i(y)\sqrt{d_i}/\sum_{i=1}^{n}q_i(x)\sqrt{d_i}K}$이다. 그러므로, $NB(p'(x),p'(y))$를 최소로 하는 최적의 정규화된 페이지 영역의 구간비 $p'(x):p'(y)=\sum_{i=1}^{n}q_i(x)\sqrt{d_i}:\sum_{i=1}^{n}q_i(y)\sqrt{d_i}$이다. 또한, 원래 페이지 영역의 구간비는 정규화된 페이지 영역의 구간비와 같으므로 최적 페이지 영역의 구간비 $p(x):p(y)=p'(x):p'(y)=\sum_{i=1}^{n}q_i(x)\sqrt{d_i}:\sum_{i=1}^{n}q_i(y)\sqrt{d_i}$이다. □

5. 결론

본 논문에서는 영역 기준 분할정책을 사용하는 다차원 파일구조를 최적으로 구성하는 방법에 관하여 논의하였다. 본 논문의 본론에서는 설명의 편의를 위하여 구성 애트리뷰트가 두 개인 이차원 파일구조에 대해서 도메인 공간의 분할 상태를 나타내는 페이지 영역들의 구간비가 질의 영역의 구간비와 같을 때, 질의 처리시에 발생되는 페이지액세스 수가 최소로 됨을 수학적으로 분석하고 증명하였다.

그러나, 구성 애트리뷰트가 두 개 이상인 다차원 파일구조에 대해서 다음과 같이 확장하여 적용할 수 있다. 즉, 주어진 n개의 N차원 질의 영역 $q_i(1)\times q_i(2)\cdots q_i(j)\cdots\times q_i(N)$ $(i=1,...,n)$에 대해서, 먼저 최적 페이지 영역의 구간비는 레코드들이 도메인 공간에서 균일하게 분포할 때는 질의 영역들의 각 축의 구간 크기를 합산한 값의 비인 $\sum_{i=1}^{n}(q_i(1)):\sum_{i=1}^{n}(q_i(2)):\cdots\sum_{i=1}^{n}(q_i(1)):\cdots\sum_{i=1}^{n}(q_i(N))$로 계산 할 수 있다. 그리고 레코드들이 도메인 공간에서 비균일하게 분포할 때는 각 질의 영역의 크기를 질의 영역에 포함되는 fp코드들의 개수가 되게 정규화하여 정규화된 n개의 질의 영역 $q_i'(1)\times q_i'(2)\cdots q_i'(j)\cdots\times q_i'(N)(i=1,...,n)$들로서 각 축의 구간 크기를 합산한 값의 비인 $\sum_{i=1}^{n}(q_i'(1)):\sum_{i=1}^{n}(q_i'(2)):\cdots\sum_{i=1}^{n}(q_i'(1)):\cdots\sum_{i=1}^{n}(q_i'(N))$로 다차원 파일구조의 최적 페이지 영역의 구간비를 계산할 수 있다.

이와같이 본 논문에서 제안하는 방법으로 다차원 파일구조의 최적 페이지 영역의 구간비를 계산하여, 이와같은 구간비의 페이지 영역들을 갖는 다차원 파일구조를 구성함으로써, 다차원 파일구조를 이용하는 많은 응용 시스템에서 성능을 매우 크게 개선할 수 있다.

참고문헌

[1] Chakrabarti, K. and Mehrotra, S. "The Hybrid Tree: An Index Structure for High Dimensional Feature Spaces," In *Proc. Intl. Conf. on Data Engineering*, pp. 440-447, 1999.

[2] Chang, J. M. and Fu, K. S, "A Dynamical Clustering Technique for Physical Database Design," In *Proc. Intl. Conf. on Management of Data*, ACM SIGMOD, pp. 188-199, Santa Monica, May 1980.

[3] Harada, L. et al., "Query Processing Method for Multi-Attribute Clustered Relations," In *Proc. Intl. Conf. on Very Large Data Bases*, pp. 59-70, Brisbane, Australi, Aug. 1990.

[4] Kim, K. H., Cha, S. K. and Kwon, K. J., "Optimizing Multidimensional Index Trees for Main Memory Access," In *Proc. of the ACM SIGMOD Conf.*, pp. 139-150, 2001.

[5] Nakamura, Y. et al., "A Balanced Hierarchical Data Structure for Multidimensional Data with Highy Efficient Dynamic Characteristics," In *Proc. Int'l Conf. on Knowledge and Data Engineering*, IEEE Trans., Vol. 5, No. 4, pp. 682-694, Aug. 1993.

[6] Nievergelt, J. et al., "The Grid File: An Adaptable, Symmetric Multikey File Structure," *ACM Trans. on Database Systems,* Vol. 9. No. 1, pp. 38-71, Mar. 1984.

[7] Orlandic, R. and Yu, B. "Implementing KDB-Trees to Support High-Dimensional Data," In *Proc. Int'l Conf. on Database Engineering & Applications Symposium,* IEEE Trans., pp. 58-67, 2001.

[8] Robinson, J. T., "The K-D-B-Tree: A Search Structure for Large Multidimensional Dynamic Indexes," In *Proc. Int'l Conf. on Management of Data*, ACM SIGMOD, pp. 10-18, Ann Arbor, Michigan, Apr. 1981.

[9] Whang, K. Y. and Krishnamurthy, R., "The Multilevel Grid File -- A Dynamic Hierarchical Multidimensional File Structure," In *Proc. Int'l Conf. on Database Systems for Advanced Applications*, pp. 449-459, Tokyo, Apr. 1991.

[10] Yu, B. et al., "KDB_{KD}-Tree: A Compact KDB-Tree Structure for Indexing Multidimensional Data," In *Proc. Intl. Conf. on Information Technology : Coding and Computing*, IEEE Trans., pp. 676-680, 2003.

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