공간 질의 최적화를 위한 힐버트 공간 순서화에 따른 공간 분할

황 환 규† · 김 현 국††

요 약

공간 질의 크기에 대한 근사치를 구하기 위해서는 입력 데이터 공간을 분할한 후 분할된 영역에 대하여 질의 결과 크기를 추정한다. 본 논문에서는 데이터 편재가 심한 공간 데이터에 대한 질의 크기 추정의 문제를 논의한다. 공간을 분할하는 기법으로 관계 데이터베이스에서 많이 사용되는 너비 균등, 높이 균등 히스토그램에 해당되는 면적 균등, 개수 균등 분할에 대한 방법을 검토하고 공간 인덱싱에 기초한 공간 분할 방법에 대해서 알아본다. 본 논문에서는 공간 순서화 기법인 힐버트 공간 채움 곡선을 이용한 공간 분할을 제안한다. 제안한 방법과 기존의 방법을 실제 데이터와 인위 데이터를 사용하여 편재된 공간 데이터에 대한 질의 결과 크기의 추정에 대한 정확도를 비교한다. 본 실험에서 힐버트 채움 곡선에 의한 공간 분할이 공간 질의 크기, 버켓수의 변화, 데이터 위치 편재도의 변화, 데이터 크기의 변화에 대해서 기존의 분할 방법보다 질의 결과 크기 추정에 대해서 우수한 성능을 보였다.

Spatial Partitioning using Hilbert Space Filling Curve for Spatial Query Optimization

Whan-Kyu Whang† · Hyun-Guk Kim††

ABSTRACT

In order to approximate the spatial query result size we partition the input rectangles into subsets and estimate the query result size based on the partitioned spatial area. In this paper we examine query result size estimation in skewed data. We examine the existing spatial partitioning techniques such as equi-area and equi-count partitioning, which are analogous to the equi-width and equi-height histograms used in relational databases, and examine the other partitioning techniques based on spatial indexing. In this paper we propose a new spatial partitioning technique based on the Hilbert space filling curve. We present a detailed experimental evaluation comparing the proposed technique and the existing techniques using synthetic as well as real-life datasets. The experiments showed that the proposed partitioning technique based on the Hilbert space filling curve achieves better query result size estimation than the existing techniques for space query size, bucket numbers, skewed data, and spatial data size.

키워드 : 공간 데이터베이스(Spatial Databases), 질의 최적화(Query Optimization), 공간 선택률 추정(Spatial Selectivity Estimation), 힐버트 공간 채움 곡선(Hilbert Space Filling Curve)

1. 서 론

공간 데이터베이스는 점, 선, 다각형, 표면과 같은 공간 데이터를 저장, 관리한다[1]. 지리정보 시스템(GIS), CAD, 컴퓨터 비전, 로보틱스, 지리학, 이미지 데이터베이스와 같이 공간 데이터 처리를 필요로 하는 응용 분야가 증가함에 따라 공간 데이터베이스의 중요성이 증가되고 있으며, 현재 상용화된 공간 데이터베이스 시스템으로는 ESPRI의 ARC/INFO[2], InterGraph의 MGE, MapInfo, Informix[3] 등이 있다. 대부분의 선도적인 데이터베이스 공급업체는 공간 데이터에 대한 지원을 제공하고 있다.

질의 최적화 모듈은 DBMS의 중요한 구성 요소이다. 일반적으로 사용자의 질의는 원하는 최종 응답을 얻기 위한 조건으로 표현된다. 요구되는 결과를 효율적으로 계산하기 위해서 여러 가지 가능한 질의 수행 계획 중 최적의 질의 수행 계획을 생성하는 것이 최적화기의 역할이다. 질의 수행 계획의 비용은 여러 가지 연산을 수행하는데 필요한 중간 결과 크기에 의하여 좌우된다. 따라서 여러 실행 전략에 대한 최소 비용 결과를 얻기 위해 중간 결과 크기를 측정할 필요가 있다. 중간 결과 크기를 측정하기 위해 선택률의 개념을 사용한다. 전체 데이터 중 검색 조건을 만족하는 데

※ 본 논문은 2001년도 강원대학교 기성회 교수 국외파견 연구 지원에 의하여 연구되었음.
† 정 회 원 : 강원대학교 전기전자정보통신공학부 교수
†† 정 회 원 : 강원대학교 대학원 정보통신공학과
논문접수 : 2003년 5월 26일, 심사완료 : 2003년 11월 12일

이터 수의 비율을 그 검색 조건의 선택률(Selectivity)이라 정의한다. 결과의 크기 측정을 실제 데이터를 통하여 수행할 경우 많은 비용이 소요됨으로 질의 결과 크기는 실제 데이터에 대한 요약된 통계 데이터로 선택률을 추정(Selectivity Estimation) 하게 된다[4]. 추정은 또한 사용자에게 질의가 실제로 수행되기 전에 대략적인 실행 시간을 제공하기 위해서도 사용된다. 질의 결과 크기의 추정값을 계산하기 위해 필요한 것은 원시 데이터의 특성을 근사치로 나타내는 요약된 데이터로 투플 수, 투플의 크기, 애트리뷰트 값 중에서 서로 다른 값의 개수 등이 사용된다. 관계 데이터베이스 시스템에서와 같이 공간 데이터베이스 시스템에서도 질의 결과의 크기를 정확하게 추정하는 모듈이 있으며 공간 데이터베이스의 중요 구성 요소이다.

관계 데이터베이스에서 질의 결과 크기를 추정하기 위한 다양한 기법들이 제시되었다. 가장 일반적인 방법으로 히스토그램[5], 샘플[6], 수학적인 분포에 의해 데이터를 모델링하는 파라메터 기법[7] 등이 있다. 각가지 기법 중에서 특히 히스토그램은 데이터베이스 시스템에서 가장 일반적으로 사용되는데(예를 들면 DB2, Oracle, Microsoft SQL Server 등) 그 이유는 실행 시간의 오버헤드가 적고, 적은 공간을 사용하며, 입력 데이터 분포가 사전에 알려질 필요가 없다는데 있다. 이 방법은 입력을 버켓이라는 작은 영역으로 분할하고 각 버켓에 대해서 입력 데이터를 근사화 한다. 질의 결과 추정은 질의를 버켓에 대한 근사치를 사용함으로써 얻어진다.

관계 데이터베이스에서 질의 크기 추정의 문제는 많은 연구가 있었지만 공간 데이터에 대한 질의 크기 추정의 문제는 그 중요성에도 불구하고 연구가 미흡한 편이다. 공간 데이터베이스의 질의 결과 크기 추정은 일반적인 질의 결과 크기 추정과 비교할 때 크게 두 가지 면에서 차이가 있다[8]. 첫째는 개개의 공간 데이터들이 서로 다른 모양과 크기를 갖는다는 것이며, 둘째는 공간상에 데이터가 비균일하게 분포할 때 빈도수는 급격하게 변하지 않고 데이터의 위치가 편재된 것(skewed)을 의미한다. 즉, 관계 데이터베이스와 달리 공간 데이터의 빈도수가 급격하게 변하지 않는다는 것은 공간 데이터가 같은 위치에서 겹치도록 존재하지 않는다는 것을 의미한다. 공간 데이터의 비균일성은 편재된 데이터에 의해서 발생한다. 다차원 데이터를 다룬 연구에서조차도[9] 편재된 데이터(즉 공간상의 데이터 위치의 편재)에 대한 고려보다 공간상에서 점에 대한 빈도수를 근사화 하는데 집중하였다. 따라서 공간 데이터 요약의 문제는 공간 데이터의 분포를 정확하게 요약할 수 있는 기술을 필요로 한다.

본 논문은 공간 데이터가 편재되었을 때 질의 결과를 추정하는 방법을 제시한다. 공간 데이터를 근사화하기 위해 전체 데이터를 버켓이라 불리는 작은 영역으로 분할하고 각 버켓에 데이터의 개수를 유지하는 히스토그램 방법을 사용한다. 본 논문에서는 다차원 인덱싱에 사용되는 공간 채움 곡선(space filling curve) 중에서 특히 데이터 클러스터링에 우수하다고 보고된[10] 힐버트(Hilbert) 공간 채움 곡선 기법을 사용하여 편재된 공간 데이터를 분할하는 방법을 제안하고 기존의 방법과 질의 결과 크기의 추정에 대한 정확성을 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 공간 데이터 분할 기법에 대해 알아본다. 3장에서 힐버트 공간 채움 곡선 기법을 사용한 공간 분할 기법을 제안하고, 4장에서는 질의 결과 크기 추정 방법을 기술하고, 5장에서 기존의 방법과 제안한 방법의 질의 결과 크기 추정의 정확성을 비교한다. 끝으로 6장에서 결론을 맺는다.

2. 기존의 공간 데이터 분할 기법

공간 데이터베이스는 다양한 모양, 서로 다른 크기의 데이터, 편재된 데이터로 구성되므로 이들을 고려하여 전체 공간을 분할 한 후, 분할된 버켓 내에서 최소 경계 사각형으로 표현된 데이터의 개수를 요약 데이터로 유지하게 된다. 모든 데이터는 분할 영역 내에 균일하게 분포되어 있음을 가정한다.

기존의 공간 데이터 분할 방법으로는 균등분할 기법과 인덱스에 기초한 분할 기법 등이 있다. 균등 분할 기법은 공간 분할에 대한 성질이 모두 같도록 분할하는 방법으로 균등 분할의 기준에 따라 분할 공간의 면적이 같도록 분할하는 면적 균등 기법과 분할 공간의 데이터 개수가 같도록 분할하는 개수 균등 기법으로 나뉜다. 인덱스 기반의 분할 기법은 공간 인덱스 구조에 의해 생성된 분할을 요약 데이터를 유지하기 위한 공간 분할로 사용하는 방법으로 본 논문에서는 가장 효율적인 공간 인덱스 구조로 알려진 R^*-트리 공간 인덱스를 공간 분할에 사용한다[11]. 최근에 데이터 밀도에 근거하여 공간 분할을 시도한 방법은[8] 최적의 데이터 분할이 NP-hard 문제가 되어 이를 해결하기 위해서 입력 영역을 수직이나 수평으로 나누는 이진 공간 분할 방법을 사용하였다. 이것도 시간 복잡도가 $O(N^{2.5})$가 되어 이를 줄이기 위해서 휴리스틱스 기법(greedy 방법)을 사용하여 지역적으로 최적의 분할을 시도하였다.

2.1 균등 분할 기법

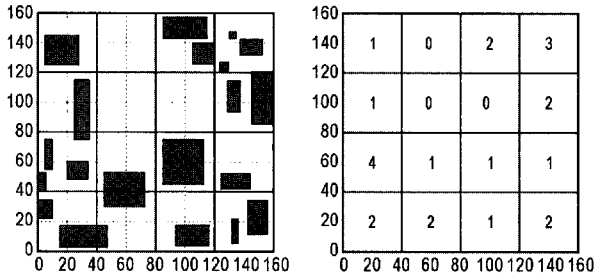
균등 분할 기법은 버켓의 면적을 균등하게 분할하는 면적 균등 분할과 버켓내의 데이터 개수를 같도록 분할하는 개수 균등 분할로 관계 데이터베이스에서 너비 균등과 높이 균등 히스토그램 방법과 유사하다[5, 12].

2.1.1 면적 균등 분할 기법

면적 균등 분할 기법은 모든 분할영역의 면적이 같아지

도록 공간을 분할하는 방법이다. 이 방법은 공간 분할 방법이 비교적 간단하지만 데이터의 분포와 무관하게 균일한 격자 형태로 공간을 분할하므로 데이터의 분포 특성을 나타내는데 한계가 있다.

(그림 1)은 면적 균등 분할 방법을 적용하여 공간 데이터를 분할한 예를 보여준다. 큰 버켓은 오차율이 커지는 가능성이 있음으로 면적을 균등하게 분할하는 것은 최악의 오류를 최소화하는 방법으로 생각할 수 있다. 분할 영역의 면적을 크게 나누면 데이터가 실제보다 더 넓은 영역에 분포한 것처럼 나타나는 문제가 발생하고 분할 영역의 면적을 작게 나누면 데이터의 개수가 실제보다 더 많이 표현되는 문제가 발생한다.

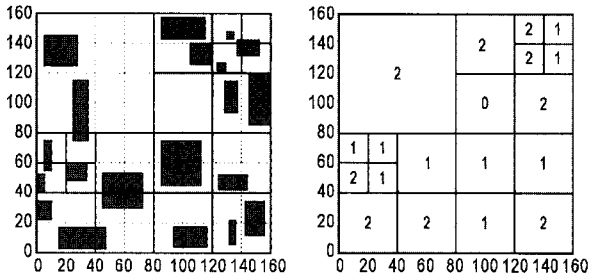


버켓 수 = 16

(그림 1) 면적 균등 분할과 요약 데이터

2.1.2 개수 균등 분할 기법

개수 균등 분할 기법은 모든 분할 영역에 존재하는 데이터의 개수가 같도록 분할하는 방법이다. 이 방법은 면적 균등 방법과는 달리 데이터의 분포를 고려하여 편재된 영역을 더 세밀히 분할한다. 하지만 편재된 데이터 영역을 중심으로 분할하다 보면 상대적으로 편재되지 않은 영역에 존재하는 데이터가 실제보다 넓은 영역으로 표현될 수 있으며 데이터의 편재가 극심해지면 버켓이 좁은 공간에 몰리면서 데이터를 중복해서 카운트하는 문제점이 있다. 또한 만일 데이터의 분포가 균일하다면 이 방법은 유용하지 못하고 버켓의 낭비를 초래할 수 있다.



(a) (b)

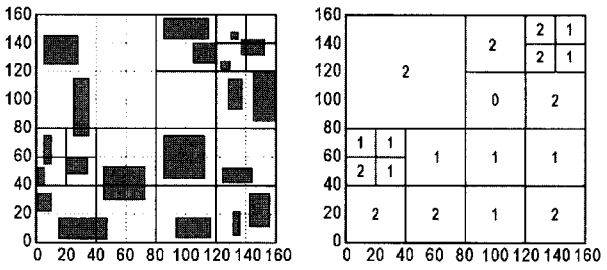
(그림 2) 개수 균등 분할과 요약 데이터

본 논문에서는 4분 트리를 사용하여 개수 균등 분할을 구현하였다. (그림 2)(a)는 모든 버켓의 데이터 개수가 가능한 한 같아지도록 편재된 영역을 더 세밀히 분할한 개수 균등 분할 결과이며 (그림 2)(b)는 분할 결과를 바탕으로 생성한 요약 데이터이다. 편재되지 않은 영역의 데이터가 넓게 분포된 것처럼 왜곡되어 나타난 것을 볼 수 있다. 또한 세밀히 분할한 영역의 데이터의 중복 카운트 현상이 발생한다.

2.2 인덱스 분할 기법

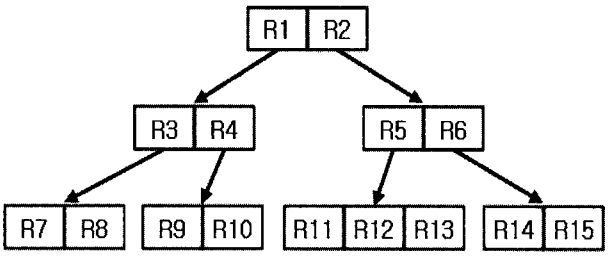
인덱스 분할 기법은 공간 인덱스 구조에 의해 생성된 분할을 요약 데이터를 유지하기 위한 공간 분할로 사용하는 방법이다. 본 논문에서는 공간 데이터 인덱스 구조로 가장 효율적이라고 알려진 R*-트리 공간 인덱스 구조를 사용한다[11]. R*-트리는 분할 공간 내에 비어 있는 공간과 분할 영역들 사이의 겹치는 영역을 최소화 한다.

(그림 3)(a)와 (그림 3)(b)는 R*-트리 인덱스 구조와 트리의 리프 노드를 최종 분할 영역으로 하는 요약 데이터를 나타낸 것으로 R*-트리 인덱스에 기초한 분할이 균등 분할에 사용한 것과는 많은 차이가 있다. 또한 (그림 4)는 (그림 3)에서 공간 분할에 사용된 R*-트리 공간 인덱스의 계층 구조를 보여준다. 분할에 사용한 엔트리의 수를 3으로 하여 인덱싱 한 결과이며 최종 분할 영역은 트리의 리프 노드인 R7~R15 영역이 된다.



(a) (b)

(그림 3) 인덱스 기반 분할과 요약 데이터 (버켓 수 = 9)



(그림 4) 분할에 사용한 공간 인덱스의 계층 구조

본 논문의 방법은 R*-트리 인덱스 구조의 특성상 리프 노드가 가리키는 버켓 영역의 중첩이 발생한다. 따라서 편재가 심한 영역일수록 더 많이 중첩된 버켓을 생성하게 되어 질의 결과 크기 추정 오차를 증가시키는 요인이 된다. 또한 R* 트리 삽입 알고리즘은 새로운 노드를 생성할 때

전체를 고려하지 않고 지역적인 모양에 의하여 결정됨으로 생성된 최종 버켓은 상당히 편재된 모양을 형성한다.

3. 제안한 공간 분할기법

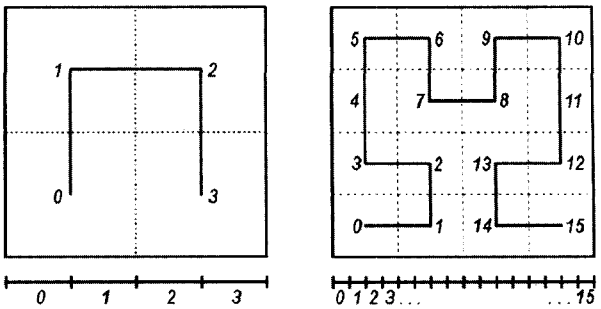
지금까지 알아본 바와 같이 기존의 분할 방법은 데이터의 분포 특성을 잘 나타내지 못한다거나 편재된 영역 혹은 편재되지 않은 영역에서의 데이터 요약 값이 왜곡되는 단점을 가지고 있다. 이를 개선하기 위해 2차원 공간을 힐버트 공간 채움(Hilbert space filling curve) 곡선 경로를 통해 진행하며 공간 분할을 수행하는 새로운 공간 분할 기법을 제안한다. 본 장에서 새로운 공간 분할 기법의 배경이 되는 힐버트 공간 채움 곡선과 제안한 공간 분할 기법에 대해 알아본다.

3.1 힐버트 공간 채움 곡선

공간 채움 곡선은 알고리즘에 따라 특정한 순서로 공간상의 모든 점을 통과하는 선이다. 곡선은 각 점을 오직 한 번만 통과하며 따라서 각 점은 시작점으로부터 곡선에 따라 유일한 거리에 있게 된다. 이러한 공간 채움 곡선은 공간을 순서화 하는 기법으로 힐버트 곡선과 z-순서 곡선 등이 있다[14, 15]. 공간 채움 곡선은 다차원 공간을 일차원 값으로 사상시켜 줌으로써 B-트리와 같이 일차원에 사용하는 인덱싱을 다차원의 인덱싱에 사용할 수 있다. 기존의 해싱 기법과는 달리 일차원 값에서 공간상의 인접성을 유지한다. 특히 공간 채움 곡선 중에서 힐버트 곡선은 다른 곡선과 비교할 때 데이터 클러스터링에 우수한 성능을 보이는 것으로 알려져 있다[10].

(그림 5)(a)에 나타난 곡선은 2차원 영역 진행 경로의 일차 순서 곡선을 나타낸다. 2차원 영역이 4개의 사분면으로 나뉜 후 일차 순서 곡선이 중심 점을 통과하며 그려진다. 인접한 두 개의 사분면은 간선(edge)을 공유하도록 사분면이 순서화 된다. (그림 5)(b)의 이차 순서 곡선은 (그림 5)(a)의 각 사분면이 재귀적으로 다시 4개의 사분면으로 나뉜다. 연속적인 사분면의 인접성이 유지되도록 이차 순서 곡선에서는 일차 순서 곡선의 방향을 달리하여 이어지는 것을 볼 수 있다.

실제 애플리케이션에서는 힐버트 곡선 생성을 위한 재귀적인 분할의 표현을 k 단계까지 진행하여 나타내는 경우 이를 순서가 k 인 힐버트 곡선이라 한다. 힐버트 곡선의 순서는 힐버트 곡선이 다차원 공간을 얼마나 세밀하게 표현하는가를 나타낸다. 순서가 k라고 할 때 n 차원 힐버트 곡선은 2^{nk}개의 공간상의 점을 통과한다. 예를 들면, (그림 5)(a)와 (그림 5)(b)는 이차원상의 순서가 각각 1과 2인 힐버트 곡선을 나타내며 각각 이차원 공간상의 점을 4개, 16개씩 통과한다.



(a) (b)

(그림 5) 2차원 힐버트 곡선의 표현

3.2 힐버트 곡선을 이용한 공간 분할 방법

앞서 기존의 공간 데이터 분할 방법으로 균등 분할 기법과 공간 인덱스 구조를 공간 분할에 적용한 방법을 살펴보았다. 균등 분할 기법은 전체 공간을 영역 중심으로 분할한 후 분할 영역에 속하는 데이터 개수를 통해 해당 영역을 요약하므로 버켓이 데이터의 분포 특성을 효율적으로 반영하지 못하는 문제점이 있다. 인덱스에 기초한 분할 기법은 버켓 영역이 겹치는 문제점과 새로운 노드의 생성이 데이터 분포를 고려하지 않고 부분적으로 이루어져 최종 버켓이 상당히 편재되는 경향이 있다.

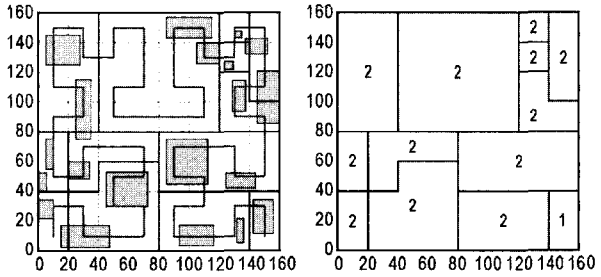
본 논문에서 제안한 방법은 힐버트 공간 채움 곡선 진행과 개수 균등 분할 기법을 혼합한 형태로 힐버트 공간 채움 곡선 경로를 따라 인접한 공간 영역을 빠짐없이 진행하다 데이터 개수가 주어진 값 이상되면 하나의 버켓을 형성하여 공간을 분할한다. 각 버켓은 대략 같은 데이터 개수를 가지며 버켓은 인접한 데이터의 밀도에 따라 나뉘게 된다. 분할은 데이터가 놓여 있는 공간을 분할하기 보다는 데이터를 분할하는 것으로 데이터의 밀집에 따라 분할이 세밀하게 이루어짐으로 공간상의 데이터 편재를 최소화한다. 제안한 힐버트 곡선을 이용한 공간 분할 알고리즘은 (Algorithm 1)에 기술하였고, 이 알고리즘을 공간 데이터에 적용하여 분할한 결과는 (그림 6)이다.

입력 : 공간 데이터
출력 : 힐버트 공간 순서화 기법을 이용하여 공간 분할 후 분할된 각 버켓에 대하여 포함된 데이터 개수를 저장
방법 : 힐버트 공간 순서화 기법을 사용하여 격자로 공간을 분할한다(격자로 분할된 영역을 쎌(Cell)이라 함). 각 쎌에 포함된 공간 데이터의 개수를 유지한 후, 주어진 값의 데이터 개수가 되도록 쎌을 합병하여 하나의 버켓을 만든다.

단계 1 : 임의의 Order k를 갖는 힐버트 곡선을 생성한다.
단계 2 : 힐버트 곡선의 리프 노드가 나타내는 영역과 겹치는 데이터의 개수를 유지한다.
단계 3 : 리프 노드를 순차적으로 스캔하면서 노드와 겹치는 데이터의 개수를 누적하여 카운트 한다.
단계 4 : 누적 데이터의 개수가 주어진 값에 도달하면 해당 영역까지를 하나의 버켓으로 저장 한다.

(알고리즘 1) 힐버트 곡선을 이용한 공간 분할 알고리즘

(그림 6)(a)는 공간 데이터의 모든 영역을 진행하는 순서가 3인 힐버트 곡선을 나타낸다. (그림 6)(b)는 힐버트 곡선의 리프 노드의 진행에 따라 공간 데이터를 스캔하면서 데이터의 개수가 주어진 값 이상이 되는 공간 영역을 하나의 버켓으로 할당하며 생성된 공간 분할 결과를 보여준다. 여기서의 예는 주어진 값이 데이터 개수가 2개가 되는 경우이고 이 값은 버켓의 개수에 의해서 정해진다.



(a) 공간 데이터 (b) 분할 결과

(그림 6) 힐버트 공간 채움 곡선을 이용한 공간 분할

제안한 방법과 기존 방법의 차이점은 요약 데이터에 위치한 버켓 영역의 모양이다. 기존의 방법은 사각형 형태의 버켓들로 요약 데이터가 이루어진 반면 제안한 방법을 통해 형성한 분할 영역은 힐버트 곡선의 진행 상태에 따라서 다양한 모양의 버켓 영역이 생성되는 것을 볼 수 있다. 이와 같은 버켓 영역은 공간 영역을 힐버트 곡선 경로에 따라 인접한 데이터를 중심으로 진행하면서 형성된 것으로 데이터의 분포 특성을 잘 반영하는 특징이 있다.

3.3 요약 데이터를 이용한 질의 결과 크기 추정

원시 공간 데이터의 공간 분할을 통해 요약 데이터가 얻어지면 특정 질의 영역에 만족하는 질의 결과 크기를 추정할 수 있다. 임의의 질의가 주어졌을 때 해당 질의에 대한 질의 결과 크기 추정값은 각 분할 영역의 요약 데이터 중 질의 영역과 겹치는 부분에 해당하는 데이터 개수의 합으로 나타낼 수 있다. 이 때 질의 영역과 겹치는 부분의 데이터 개수는 분할 영역이 유지하고 있는 요약 데이터의 개수를 질의 영역에 겹치는 비율로 곱하여 나타낸다. 이는 공간 분할을 통해 얻어진 요약 데이터의 개수가 분할 영역 내에 균일하게 분포한다는 가정에 기초한 계산 방법이다. 질의 결과 크기 추정 공식을 정리하면 다음과 같다.

$$S(Q) = \sum_{i=1}^{k} n_i \times r_{i_{Overlap}}$$

S(Q) : 질의 Q 에 대한 질의 결과 크기 추정값
k : 질의 영역과 겹치는 분할 영역의 개수
n_i : i 번째 분할 영역 내 요약데이터의 개수($1 \le i \le k$)
$ri_{Overlap}$: i 번째 분할 영역과 질의 영역이 겹치는 비율

4. 성능 평가

본 장에서는 다양한 공간 분할 기법의 질의 결과 크기 추정 성능을 알아본다. 실험은 주어진 실제 데이터와 인위 데이터를 기준으로 수행하였으며 각각의 공간 분할 방법을 통해 생성된 요약데이터의 정확성을 평가 한다. 다양한 공간 분할 방법의 성능을 비교하기 위하여 다음의 상대 오차율 공식을 사용한다.

$$E(Q) = \frac{|r-e|}{r} \times 100$$

E(Q) : 질의 Q에 대한 상대 오차율(%)
r : 실제 질의 결과 크기
e : 추정된 질의 결과 크기

앞서 설명한 질의 결과 크기 추정 공식에 의한 질의 결과 크기 추정값과 실제 질의 결과 크기와의 상대 오차율을 성능 평가의 기준으로 삼는다. 실제 질의 결과 크기란 분할 이전의 데이터 중에서 질의 영역에 속하는 모든 데이터의 개수를 의미한다.

4.1 실험 환경

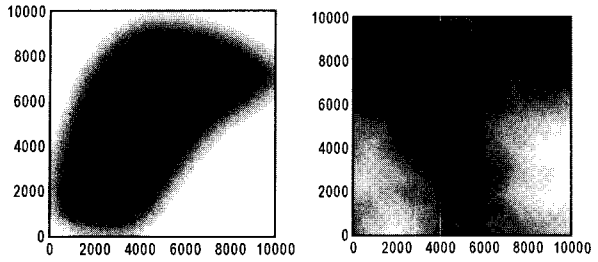
실험에 사용되는 데이터 집합과 질의 집합은 아래와 같으며 각각의 실험은 공간 분할 방법을 제외한 모든 조건이 동일한 상태에서 진행되도록 하였다. 실험은 Sun Ultrasparc Workstation에서 C로 구현하였다.

4.1.1 데이터 집합

다양한 공간 분할 기법의 성능 평가를 위해 <표 1>에서와 같이 설명된 실제 데이터와 인위 데이터 집합을 사용하였다. 실제 데이터는 일반적으로 공간 데이터베이스 연구에서 많이 사용되는 Long Beach Data[13]를 사용하였다. 인위 데이터는 데이터의 위치 편재도, 크기, 개수 등을 달리하여 직접 생성하여 실험 하였다. 위치의 편재도는 Zipf 분포[14]를 2차원에 적용하여 나타냈다. Zipf 분포의 인자인 Z 값은 데이터 위치의 편재 정도를 수치로 표현한다. 본 논문에서 사용한 Z 값의 범위는 0부터 2까지로 Z 값이 0이라는 것은 균일한 분포를 의미하며 값이 커질수록 데이터의 편재가 심해지는 것을 의미한다. <표 1>은 실험에서 사용된 데이터 집합에 대한 세부 명세이며 (그림 7)는 실험에 사용된 데이터의 대략적인 분포 형태를 보여준다.

〈표 1〉 실험에 사용된 데이터 집합

구 분	실제 데이터	인위 데이터
공간영역크기	10000×10000	10000×10000
데이터 개수	53145	50000
위치 편재도	실제 분포	Z = 0, 0.25, 0.5, 0.75, 1, 2
분 포 형 태	실제 분포	Random



(a) Long Beach (b) Zipf Random

(그림 7) 실험에 사용된 공간 데이터의 분포 형태

4.1.2 질의 집합

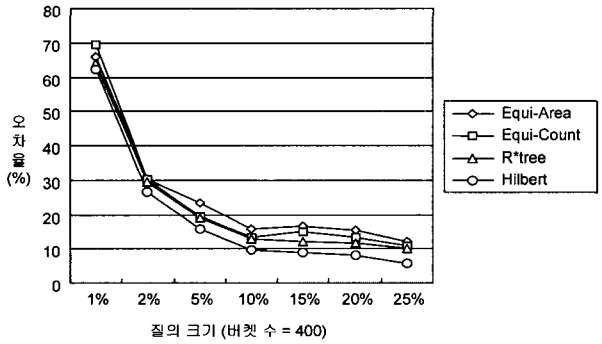
실험에 사용된 질의 집합은 무작위로 생성한 100개의 사각 영역으로 이루어진다. 질의 영역의 위치는 임의로 선택된 공간 데이터의 중점을 기준으로 정해지며 그 크기는 전체 영역 너비와 높이의 1%~25%(전체 공간 면적의 0.01%~6.25%)가 되도록 생성하였다.

4.2 실험 결과

본 장에서 다양한 공간 분할 기법의 질의 결과 크기 추정 정확성을 비교한다. 실험 결과로 나타난 오차율은 앞서 다루었던 상대 오차율을 의미한다. 또한 실험 결과의 표현을 간략히 하기 위해 면적 균등 분할 기법은 Equi-Area, 개수 균등 분할 기법은 Equi-Count, 인덱스 기반 분할 기법은 R*-tree로 표현하며 본 논문이 제안한 힐버트 공간 채움 곡선에 의한 분할 기법은 Hilbert로 각각 표현한다.

4.2.1 질의 크기의 변화에 따른 성능

본 실험은 질의 크기의 변화에 따른 각 공간 분할 방법의 성능을 보여준다. (그림 8)은 Long Beach 데이터를 400개의 버켓으로 고정시켜 분할한 후 질의 크기 변화에 대한 상대 오차율을 구한 결과이다.

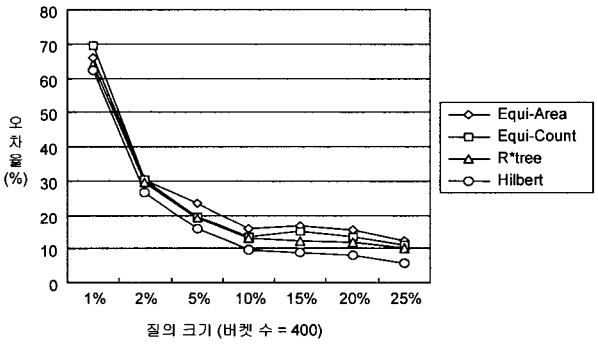


(그림 8) 질의 크기 변화에 따른 성능

일반적으로 질의 크기가 커짐에 따라 오차율이 줄어드는 경향을 보인다. 버켓은 버켓이 포함하는 영역의 정확한 데이터 개수를 유지하기 때문에 질의 영역에 완전히 포함되는 버켓에서는 추정 오차가 발생하지 않으며 오차는 질의 영역에 부분적으로 겹치는 버켓에서만 발생한다. 그러므로 질의 크기가 클수록 질의 영역과 부분적으로 겹치는 버켓의 수가 감소하기 때문에 오차율 감소를 가져오게 된다. 면적 균등 분할, 개수 균등 분할, R*-트리 분할 방법은 유사한 오차율을 보이고 질의 크기가 5% 이상에서 힐버트 분할 방법이 15%에서 40%의 성능 향상이 되는 것을 볼 수 있다.

4.2.2 버켓 수의 변화에 따른 성능

본 실험은 버켓 수의 변화에 따른 각 공간 분할 방법의 성능을 보여준다. (그림 9)는 Long Beach 데이터를 사용하여 버켓수를 100에서 500으로 변화시킬 때 평균 상대 오차율을 구한 그래프이다. 데이터의 요약을 위해 더 많은 수의 버켓을 사용한다는 것은 그만큼 데이터의 특성을 잘 반영하는 것이므로 대체로 버켓 수가 많아짐에 따라 오차율이 줄어드는 경향을 보인다. 하지만 실험 결과 중 면적 균등 분할은 버켓 수가 증가함에 따라 오차율 또한 증가하는 경향을 보이는데 그 이유는 많은 버켓이 생성되면서 두 개 이상의 버켓에 중복하여 기록되는 데이터가 많이 발생하기 때문이다. 개수 균등 분할과 R*-트리는 유사한 오차율을 보인다.



(그림 9) 버켓 수 변화에 따른 성능

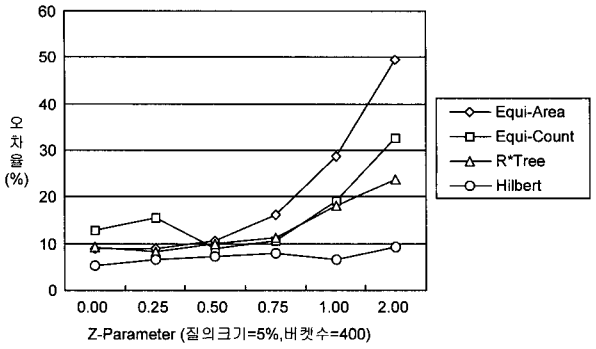
R*-트리 분할 방법에서 버켓의 개수를 조정하는 어려움이 있다. 이 문제는 주어진 버켓 수를 초과하지 않으면서 원하는 버켓 수에 근접하도록 트리의 자식 노드 수를 조정하였다.

4.2.3 위치 편재도의 변화에 따른 성능

본 실험은 위치 편재도의 변화에 따른 각 공간 분할 방법의 성능을 보여준다. (그림 10)은 Zipf 분포의 임의 데이터를 공간 분할을 통해 요약 데이터로 나타낸 후 상대 오차율을 구한 결과이다. 질의 크기가 5%, 버켓 수가 400개인 경우의 실험 결과이다.

면적 균등 방법의 오차율은 Z 값이 0.5 ~ 0.75를 기점으로 급격히 증가한다. 면적 균등 방법은 모든 영역을 균일한 격자 형태로 분할하여 나타내기 때문에 위치 편재도가 심

해질수록 낮은 정확성을 보인다.

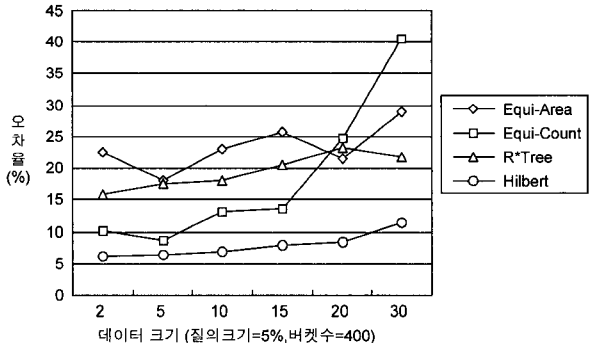


(그림 10) 위치 편재도의 변화에 따른 성능

개수 균등 방법과 R*-트리 방법은 유사한 오차율을 보인다. 힐버트 분할 방법 이외는 Z값이 커질수록 (즉 데이터의 위치 편재도가 심한 경우) 심한 오차율을 보였으나, 힐버트 분할 방법에서는 Z값의 변화에 거의 관계없이 일정한 오차율(5%~10%)을 보였다.

4.2.4 데이터 크기의 변화 따른 성능

본 실험은 데이터 크기의 변화에 따른 각 공간 분할 방법의 성능을 보여준다. (그림 11)은 Zipf 분포의 임의 데이터를 공간 분할을 통해 요약 데이터로 나타낸 후 상대 오차율을 구한 결과이다. 질의 크기가 5%, 버켓 수가 400개인 경우의 실험 결과이며 임의 데이터 생성시 Z값은 0.75를 사용하였다.



(그림 11) 데이터 크기의 변화에 따른 성능

데이터 크기의 변화에 대해 개수 균등 방법이 가장 큰 성능 변화를 보였다. 개수 균등 방법은 데이터 크기가 커짐에 따라 오차율이 급격히 증가한다. 이는 데이터의 크기가 클수록 버켓에 중복 기록될 가능성이 많기 때문이다. 개수 균등 분할, 면적 균등 분할, R*-트리 분할 방법은 10%에서 25%의 오차율을 보였으나 힐버트 분할 방법은 오차율이 5%에서 10% 사이로 데이터 크기 변화에 크게 영향 받지 않고 안정적인 오차율을 보였다.

위의 모든 실험에서 실제 데이터와 인위 데이터를 사용하여 질의 크기, 버켓수, 위치 편재도의 변화, 데이터 크기의 변화에 대하여 질의 결과 추정에 대한 정확도를 비교해 본 결과 제안한 방법이 가장 우수한 성능을 보였다.

5. 결 론

공간 선택률은 공간 데이터베이스의 질의 처리에 중요한 요소이다. 공간 데이터베이스의 관심이 증가함에도 불구하고 공간 선택률 추정에 대한 정확한 기법을 제공하는 방법에 대한 연구가 미진하였다. 공간 데이터는 관계 데이터와 많은 차이가 있음으로 관계 데이터베이스에서 사용한 기법을 잘 적용할 수 없는 어려움이 있다. 공간 데이터베이스에서는 질의 최적화를 위해 실제 공간 데이터의 특성을 근접하게 나타내는 요약 데이터를 생성하게 되며, 그 결과를 통해 질의 결과의 크기를 추정한다. 요약 데이터 생성을 위해 기존에 연구된 공간 분할 방법으로는 균등 분할 기법, 공간 인덱스에 기초한 분할 기법 등이 있다.

본 논문은 기존 방법의 공간 분할로 인해 발생하는 문제점들을 해결하기 위해 힐버트 공간 채움 곡선에 개수 균등 분할 기법을 적용한 새로운 공간 분할 방법을 제안하고 실험을 통해 기존의 방법과 질의 결과 크기 추정의 정확성을 비교하였다. 제안한 방법이 기존 방법보다 질의 크기, 버켓 수, 위치 편재도, 데이터 크기 변화에 대하여 우수한 성능을 보였다. 제안한 방법이 힐버트 곡선의 진행에 따라 공간 영역을 진행하며 데이터를 중심으로 각각의 버켓 영역을 형성해 나가기 때문에 인접한 데이터의 분포 특성을 더 잘 표현하는 특징을 보여 기존의 방법에 비해 안정적인 성능과 높은 추정 정확성을 보여 주었다. 향후 논문의 발전 방향으로 공간 데이터의 갱신이 이루어질 때 공간 분할을 전체적으로 수행하는 것은 비용이 많이 소요됨으로 이미 이루어진 분할에 최소한의 영향을 주는 효율적인 분할 방법이 필요하며 이에 대한 연구가 요구된다.

참 고 문 헌

[1] Guting, R. H., "An Introduction to Spatial Database Systems," The VLDB Journal, Vol.3, No.4, pp.357-400, October, 1994. .

[2] ARC/INFO, "Understaning GIS - the ARC/INFO Method," ARC/INFO, 1993.

[3] Ubell, M., "The Mantage Extensible Datablade Architecture," Proc. SIGMOD Intl. Conf. on Management of Data, 1994.

[4] Selinger, P., M. M. Astrahan, D. D. Chamberin, R. A. Lorie, T.G. Price, "Access Path Selection in a Relational Database Mangement System," Proc. SIGMOD Intl. Conf. on Management of Data, pp.23-34, 1979.

[5] Poosala, V., Y. Ioannidis, P. Haas and E. Shekida, "Improved Histogram for Selectivity Estimation of Range Predicates," Proc. SIGMOD Intl. Conf. on Management of Data, pp.294-305, 1996.

[6] Lipton, R. J., J. F. Naughton and D. A. Schneider, "Practical Selectivity Estimation through Adaptive Sampling," Proc. SIGMOD Intl. Conf. on Management of Data, pp. 1-11, 1990.

[7] Chen, C. M. and N. Roussopoulos, "Adaptive Selectivity Estimation using Query Feedback," Proc. SIGMOD Intl. Conf. on Management of Data, pp.161-172, 1994.

[8] Acharya, S., V. Poosala and S. Ramaswamy, "Selectivity Estimation in Spatial Databases," Proc. SIGMOD Intl. Conf. on Management of Data, 1999.

[9] Poosala, V. and Y. Ioannidis, "Selectivity Estimation without the Attribute Value Independence Assumption," Proc. SIGMOD Intl. Conf. on Management of Data, 1997.

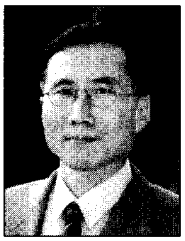
[10] Faloutsos, C. and S. Roseman, "Fractals for Secondary Key Retrieval," Proc. SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp.247-252, 1989.

[11] Beckman, N., H-P Kriegel, R. Schneider and B. Seeger, "The R*-Trees : An Efficient and Robust Access Method for Points and Rectangles," Proc. SIGMOD Intl. Conf. on Management of Data, pp.322-331, 1990.

[12] Piatetsky-Shapiro, G., and C. Connell, "Accurate Estimation of the Number of Tuples Satisfying a Condition," Proc. SIGMOD Intl. Conf. on Management of Data, 1984.

[13] Tiger/line files(tm), 1992 Technical Documentation, Technical Report, U. S. Bureau of the Census, 1992.

[14] Zipf, G. K, "Human behavior and the principle of least effort," Addison-Wesley, 1949.



황 환 규

e-mail : wkwhang@kangwon.ac.kr

1976년 서울대학교 공과대학(학사)

1987년 플로리다대학교 전기공학과(석사)

1991년 플로리다대학교 전기공학과(박사)

1992년~1994년 한국전자통신연구소

1994년~현재 강원대학교 전기전자정보통신공학부 교수

관심분야 : DBMS, 공간 데이터베이스, 데이터마이닝, 데이터웨어하우징

김 현 국

e-mail : johnk21@orgio.net

2001년 강원대학교 정보통신공학과(학사)

2003년 강원대학교 대학원 컴퓨터정보통신공학과(석사)

2003년~현재 (주)웅진에스티 SM사업2부 근무

관심분야 : 데이터베이스시스템, 공간 데이터베이스, 데이터웨어하우징