질의 결과 크기 추정을 위한 효과적인 공간 분할 기법

김현국[0] 김학자 황환규
강원대학교 컴퓨터정보통신공학과
(johnk[0],lucia)@mirae.kangwon.ac.kr wkwhang@kangwon.ac.kr

Effective Spatial Partitioning Technique for Query Result Size Estimation

Hyun-Guk Kim[0] Hak-Ja Kim Whan-Kyu Whang
Dept. of Computer Information & Telecom., Kangwon National University

요 약

공간 데이터베이스의 규모는 매우 방대하여 질의 처리에 많은 비용이 발생한다. 따라서 효율적인 질의 처리를 위해서는 질의 수행 결과의 예측이 필요하다. 이를 위해 실제 공간 데이터의 특성을 근접하게 나타내는 요약 데이터를 생성하여 그 결과를 통해 질의 결과의 크기를 추정하게 된다. 기존의 공간 데이터 요약 기법으로는 면적 균등 분할 기법, 개수 균등 분할 기법, 인덱스 분할 기법 등이 있다. 본 논문에서는 기존에 연구된 다양한 분할 기법에 대해 알아보고, 힐버트 공간 채움 곡선 방법에 개수 균등 분할 기법을 적용시킨 새로운 공간 분할 방법을 제안하여 기존의 방법과 새로운 방법의 성능을 비교한다.

1. 서 론

공간 데이터베이스의 질의 최적자는 다양한 질의 수행 계획들 중 최소의 접근 비용을 갖는 효율적인 계획을 선택하여 질의 처리에 반영한다[1]. 공간 데이터베이스는 그 규모가 매우 크기 때문에 질의 최적자는 공간 데이터의 분포, 크기, 모양 등을 고려하여 실제 데이터 분포를 최대한 근사화할 수 있는 요약 데이터를 유지하여 질의 결과 크기를 사전에 추정한다.

요약 데이터를 생성하기 위한 방법으로 전체 데이터를 버켓이라 불리는 작은 영역으로 분할하고 각 버켓에 데이터의 개수를 유지하는 히스토그램 방법이 많이 사용된다[1][2]. 요약 데이터 생성을 위해 기존에 연구된 방법으로 균등 분할 기법, 공간 인덱스에 기초한 분할 기법 등이 있다[3]. 균등 분할 기법은 버켓의 면적을 같도록 분할하는 면적 균등 기법과 버켓의 데이터 개수가 같도록 분할하는 개수 균등 분할 기법이 있다. 또한 인덱스 분할 기법은 공간 인덱스 구조에 의해 생성된 분할을 요약 데이터를 유지하기 위한 공간 분할로 사용하는 방법이다.

본 논문에서는 수학 분야에서 널리 논의 되어 온 힐버트 공간 채움 곡선 방법에 개수 균등 분할 기법을 적용한 새로운 공간 데이터 분할 방법을 제안하고 기존의 방법과 질의 결과 크기 추정의 정확성을 비교한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 공간 데이터 분할 기법에 대해 알아보고, 3장에서는 본 논문에서 제안한 공간 분할 기법에 대해서 기술 한다. 4장에서 기존의 방법과 제안한 방법의 성능을 실험 결과를 통해 비교한 후 5장에서 결론을 맺는다.

2. 기존의 공간 데이터 분할 기법

공간 데이터베이스는 다양한 모양, 서로 다른 크기의 데이터로 이루어지기 때문에 전체 공간을 분할한 후 분할된 버켓 내에서 최소 경계 사각형(MBR, Minimum Bounding Rectangle)으로 표현된 데이터의 개수를 요약 데이터로 유지하게 된다[3]. 또한 모든 데이터는 분할 영역 내에 균일하게 분포되어 있음을 가정한다.

공간 분할 시 실제 공간 데이터의 특성을 최대한 유사하게 나타낼 수 있도록 요약 데이터를 생성하기 위하여 MBR의 개수와 크기, 위치 등을 고려하여 분할한다. 그림1(a)는 공간 데이터의 간단한 예이다.

2.1. 균등 분할 기법

공간 영역을 나누는 버켓의 면적이나 버켓에 존재하는 데이터의 개수가 같도록 전체 공간을 분할하는 방법이다.

2.1.1. 면적 균등 분할 기법

모든 분할 영역의 면적이 같아지도록 데이터의 분포와 무관하게 균일한 격자 형태로 버켓을 생성한다. 그림1(b)는 주어진 공간 데이터를 같은 면적을 갖는 64개의 버켓으로 분할한 예를 보여준다. 버켓 내부의 숫자는 해당 영역에 존재하는 MBR 개수를 나타낸다.

2.1.2. 개수 균등 분할 기법

모든 분할 영역에 존재하는 데이터의 개수가 임계값 이내에서 같아지도록 분할 한다. 면적 균등 방법과는 달리 데이터의 분포를 고려하여 편재된 영역을 더 세밀히 분할 한다. 그림1(c)는 MBR 개수가 4개 이상인 영역만을 분할하여 만들어진 분할 결과이다.

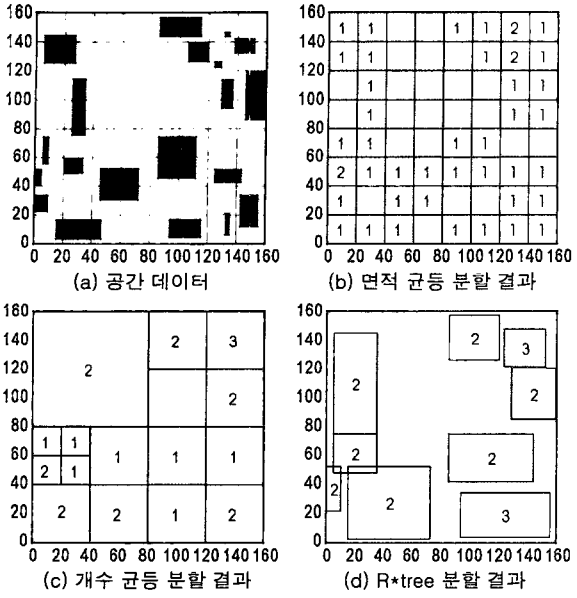

그림1. 공간 데이터와 다양한 공간 분할 결과

2.2. 인덱스 분할 기법

공간 인덱스 구조에 의한 공간 분할 방법이다. 가장 효율적인 공간 인덱스 구조인 R*tree 공간 인덱스 구조를 중심으로 논의 한다. R*tree는 분할 공간 내에 비어 있는 공간과 분할 영역들 사이의 겹치는 영역을 최소화 하는 형태로 분할을 수행 한다[4]. 그림1(d)는 R*tree 공간 인덱스 구조를 통해 리프 노드의 개수를 3개 이내가 되도록 분할한 결과이다.

3. 제안한 공간 분할 기법

기존의 균등 분할 방법과 인덱스 분할 방법은 편재된 데이터의 특성을 상대적으로 잘 표현하지 못하는 단점을 가지고 있다. 이를 개선하기 위해 본 논문에서는 2차원 데이터를 힐버트 공간 채움 곡선 경로를 통해 스캔하며 공간 분할을 수행하는 새로운 공간 분할 기법을 제안 한다.

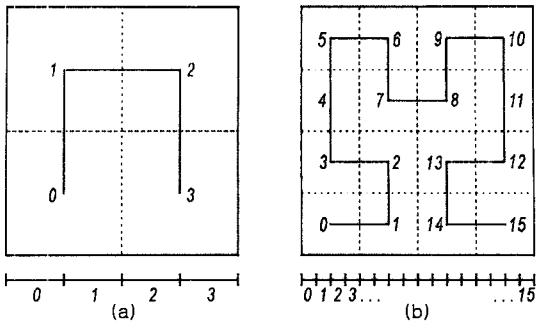

그림2. 2차원 힐버트 곡선의 표현

3.1. 힐버트 공간 채움 곡선

힐버트 곡선은 다차원 공간을 특정한 패턴[5]을 그리며 빠짐 없이 스캔 한다. 그림2에서 보는 바와 같이 힐버트 곡선을 통해 진행된 경로는 끊어짐 없이 모든 2차원 공간 영역을 재귀적으로 분할하여 1차원과 같이 스캔 함을 알 수 있다. 힐버트 곡선의 Order는 힐버트 곡선이 다차원 데이터를 얼마나 세밀하게 진행하는가를 나타낸다. Order가 k라고 할 때 n차원 힐버트 곡선은 2^{nk}개 만큼의 다차원 영역을 진행한다[6]. 예를 들면, 그림2의 (a)와 (b)는 Order가 각각 1과 2인 힐버트 곡선을 나타낸다.

3.2. 트리 구조 표현

힐버트 곡선이 재귀적으로 다차원 공간 영역을 진행해 가는 방법을 트리 구조를 사용하여 잘 표현 할 수 있다[5][6]. 그림3은 기본 곡선을 형성하고 있는 각 노드와 분할된 자식 노드의 계층 구조를 나타낸다. 각 노드는 해당 공간 영역의 좌표, 힐버트 곡선의 진행 순서 등의 정보를 유지하며 이 정보를 기준으로 자식 노드를 분할한다[5].

3.3. 힐버트 곡선을 이용한 공간 분할 방법

인접한 영역을 빠짐없이 진행해 나가는 힐버트 곡선의 특징을 공간 분할에 적용한 기본 알고리즘은 다음과 같다.

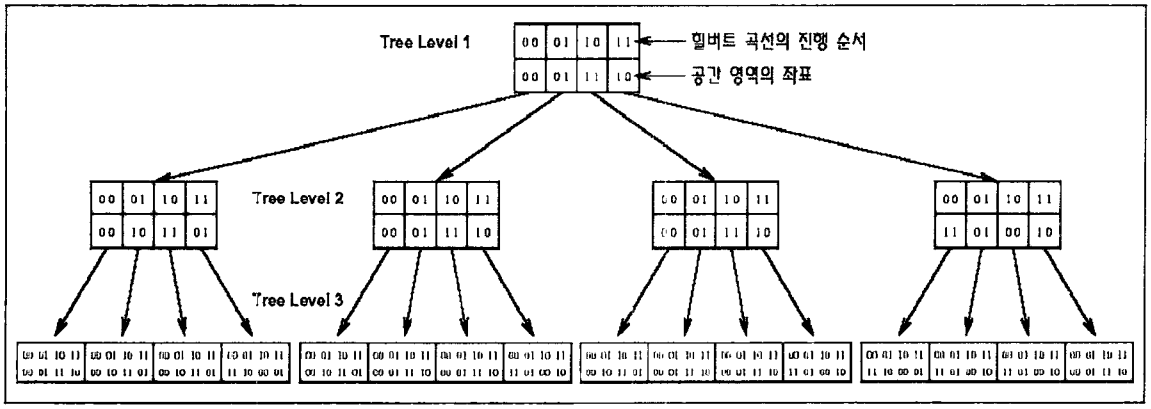

그림3. 2차원 힐버트 곡선의 트리 구조 표현 (Order = 3)

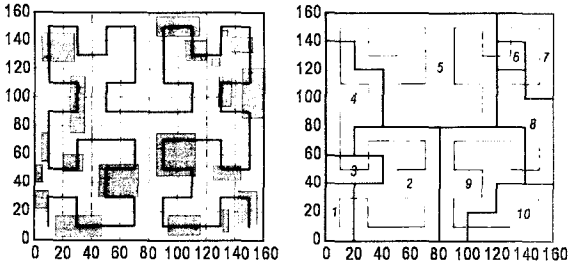

그림4. 힐버트 공간 채움 곡선을 이용한 공간 분할

단계1. 임의의 Order k를 갖는 힐버트 곡선을 생성 한다.
단계2. 힐버트 곡선의 리프 노드가 나타내는 영역과 겹치는 데이터의 개수를 유지 한다.
단계3. 리프 노드를 순차적으로 스캔하면서 노드와 겹치는 데이터의 개수를 누적하여 카운트 한다.
단계4. 누적 데이터의 개수가 임계 값에 도달하면 해당 영역까지를 하나의 버켓으로 저장 한다.

그림4(a)는 공간 데이터의 모든 영역을 진행하는 Order가 3인 힐버트 곡선을 나타낸다. 그림4(b)는 힐버트 곡선의 리프 노드의 진행에 따라 공간 데이터를 스캔하면서 만들어진 공간 분할 결과를 보여준다.(임계값 = 2) 힐버트 곡선을 통해 생성된 버켓은 데이터의 개수에 의해 분할되므로 편재된 데이터를 분할하는데 좋은 성능을 기대할 수 있다.

4. 성능 평가

본 장에서는 기존의 방법과 새 방법의 성능을 실험을 통해 비교하여 제안된 방법의 성능 향상을 보이고자 한다.

4.1. 실험 방법

데이터는 일반적으로 공간 데이터베이스에 많이 사용되는 Long Beach Data[7]를 사용하였고, 각각의 방법에 대해서 질의 크기와 버켓 수를 변화 시키면서 진행하였다. 또한 힐버트 곡선 Order를 7로 하여 힐버트 곡선을 생성하였다.

4.2. 실험 결과

그림5에서 보는 바와 같이 버켓 수를 고정시킨 상태에서 질의 크기를 변화시킬 때 질의 크기가 커질수록 전체적인 오차율이 줄어드는 것을 확인 할 수 있다. 그 이유는 질의 크기가 커질수록 질의 영역과 부분적으로 겹치는 분할 영역이 감소하기 때문이다.

그림6의 경우는 질의 크기를 고정시키고 버켓 수를 변화시키면서 진행한 결과로서 버켓 수가 많아 질수록 5%의 비교적 작은 질의에서 전반적으로 좋은 성능을 보인다.

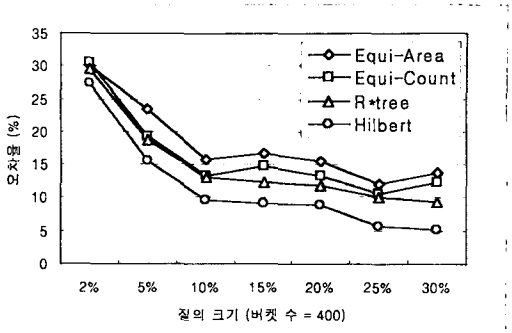

그림5. 질의 크기 변화에 따른 성능

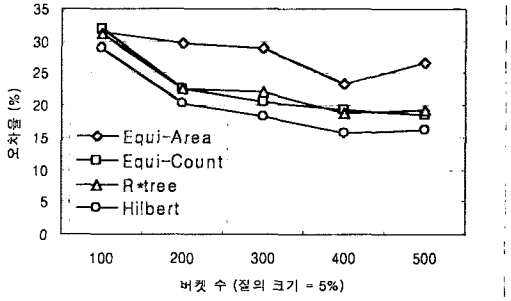

그림6. 버켓 수 변화에 따른 성능

5. 결 론

공간 데이터베이스는 데이터 규모의 특성상 효율적인 질의 결과 크기의 추정이 필수적이다. 본 논문에서는 힐버트 곡선을 사용한 효율적인 공간 분할 방법을 제안하였다. 힐버트 곡선 공간 분할 방법은 데이터가 존재하는 영역을 세밀히 스캔하여 요약데이터를 작성하게 되므로 기존의 방법에 비해 데이터의 특성을 잘 반영하여 실험 전반에 걸쳐 안정적인 성능 분포와 성능 향상을 보여 주었다.

6. 참고문헌

[1] Yannis E. Ioannidis. "Query Optimization", SIGMOD, 1996.
[2] Viswanath Poosala, Yannis E. Ioannidis, Pete J. Haas, and Eugene J. Shekita, "Improved Histogram for Selectivity Estimation of Range Predicates", SIGMOD, 1996
[3] Swarup Acharya, Viswanath Poosala, and Sridhar Ramaswamy. "Selectivity Estimation in Spatial Databases", SIGMOD, 1999.
[4] N. Beckmann, H. P. Kriegiel, R. Schneider, & B. Seeger. "The R*tree: An Efficient and Robust Access Method for Points and Rectangles", SIGMOD, 1990
[5] J. K. Lawder & P. J. H. King. "Using Space-filling Curves for Multi-dimensional Indexing", BNCOD 17, 2000.
[6] J. K. Lawder & P. J. H. King. "Querying Multi-dimensional Data Indexed Using the Hilbert Space-Filling Curve", SIGMOD 2001
[7] Tiger / line files TM, 1992 technical documentation. Technical report, 1992