공간 데이터 웨어하우스에서 분포 지역 질의 처리를 위한 확장된 큐브 트리 기법

°최준호*, 유병섭*, 박순영*, 배해영*
인하대학교 컴퓨터 정보공학과
e-mail : jhChoi@dblab.inha.ac.kr

The Extended Cube Tree for Distribution Area Query Processing in Spatial Data Warehouses

°Jun-Ho Choi*, Byeong-Seob You*, Soon-Young Park*, Hae-Young Bae*
*Dept. of Computer Science & Information Engineering, Inha Univ

요 약

최근 원격 탐사 시스템 등이 발전함에 따라 축적된 공간 데이터의 양이 증가했고 이를 공간 데이터 웨어하우스 분야에서 의사 결정에 활용하는 방안이 중요한 이슈가 되고 있다. 기존의 활용 방법은 주어진 영역을 기준으로 공간 범위-집계를 검색하는 형태였지만, 최근 특정 성향 분석을 위해 분포 질의를 요청하고 그 결과 지역에 대한 공간 분석을 통한 의사결정의 필요성이 대두되었다. 하지만 기존의 처리 방법으로 비공간 질의를 처리하기 위해서는 모든 데이터를 검색해야 하므로 분포 질의를 처리하기 위한 비용이 증가하게 된다.

본 논문에서는 분포 지역 질의 처리를 위한 확장된 큐브 트리 기법을 제안한다. 제안하는 기법은 분석하고자 하는 사실 테이블의 비공간 속성을 큐브 트리의 키로 사용하고, 이 속성과 관련된 공간 데이터의 포인터 집합을 관리한다. 본 논문의 제안 기법을 공간 데이터 웨어하우스에 적용함으로써 비공간 속성 질의를 통해 공간 객체를 결과로 요청하는 형태의 질의를 지원할 수 있게 되며 사실 컬럼을 계층화시킴으로서 사용자에게 좀 더 다각적인 분석을 지원할 수 있다.

1. 서 론

데이터 웨어하우스는 의사 결정을 효과적으로 지원하기 위해 수년간 축적된 데이터를 주제별로 통합하여 저장해 놓은 데이터 저장소이며, OLAP(On-Line Analytical Processing)은 의사 결정을 위해 사용자가 데이터 웨어하우스에 접근하여 다차원적 정보를 분석 할 수 있도록 하는 기술이다[1]. 근래 위성 원격 검침 시스템(satellite telemetry system)이나 원격 탐사 시스템(remote sensing system)등의 발전과 대중적인 사용에 따라 막대한 양의 공간 데이터가 공간 데이터베이스 및 지리 정보 시스템에 축적되어져 왔으며, 상점 개설 시의 부지 선정, 지점들의 판매량 분석 또는 기후 분석 등의 의사 결정에 축적된 공간 데이터를 활용하는 방법이 데이터 웨어하우스 분야에서 중요한 이슈가 되고 있다[2]. 이와 같이 질의의 결과 값으로 분포지역을 요청하는 범위 질의를 본 논문에서는 분포 지역 질의라고 한다.

공간 데이터 웨어하우스의 OLAP 연산은 OLTP(On-Line Transactional Processing) 시스템에서의 질의응답 속도에 비해 수십 배 이상 오래 걸린다. 이런 문제점을 해결하기위해 공간 데이터의 효율적인 OLAP 연산에 관한 많은 연구가 진행되고 있다. 하지만 기존 연구의 진행 방향은 주어진 영역을 기준으로 공간 범위-집계를 검색하는 질의의 형태이며, 주로 R-Tree 색인 기반의 질의 처리에 중점을 두고 있다[3][4]. 그러나 특정 성향 분석을 위해 분포 질의를 요청하고, 그 결과 지역에 대한 공간 분석을 통해 의사결정을 내려야 하는 상황이 빈번해지면서 비공간 속성을 통해 분포 지역과 같은 공간 속성의 결과를 분석할 수 있는 인덱스가 필요하다. 분포 질의 결과는 MBR을 기준으로 관리될 수 없기 때문에 R-Tree[5] 색인 기법으로 관리되지 않는다. R-Tree 색인 기반의 시스템에서 이런 형태의 질의를 처리하기 위해서는 모든 데이터를 검색해야 하므로 처리 비용이 증가한다.

본 논문1)에서는 비공간 속성 질의를 공간 속성 결과로 변환하는 형태의 질의를 효율적으로 처리하기 위해 확장된 큐브 트리 (Cube-Tree) 색인 기법을 제안한다. 제안기법은 분석하고자 하는 사실 테이블의 비공간 속성을 큐브 트리의 키로 사용한다. 그리고 단말 노드에 동일한 비공간 속성 값을 가진 공간 데이터에 대한 포인터 집합을 관리함으로서 비공간 속성 질의를 공간 속성 결과로 변환한다. 제안 기법을 공간 데이터 웨어하우스 시스템에 구축함으로써 비공간 속성 질의를 통해 공간 객체를 결과로 요청하는 형태의 질의를 노드의 탐색 비용을 줄이면서 지원할 수 있게 된다. 또한 큐브 트리를 기반으로 하기 때문에 OLAP 연산에 필수적인 계층(hierarchy) 구조를 지원할 수 있으며 사실 컬럼을 계층화시킴으로서 사용자에게 좀 더 다각적인 분석을 지원할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 기반 환경인 공간 데이터 웨어하우스와 큐브 트리 색인 기반의 질의 처리 기법에 대해서 설명한다. 3장에서는 본 논문에서 제안하는 확장된 큐브 트리의 전체적인 구조 및 자료구조, 질의처리 알고리즘에 대해서 살펴보고 4장에서는 성능평가를 수행한다. 마지막으로 5장에서는 결론 및 향후 연구를 하고 마친다.

2. 관련 연구

이 장에서는 본 연구의 기반 시스템인 공간 데이터 웨어하우스, 공간 OLAP 연산과 OLAP 연산 지원을 위한 큐브 트리의 구조에 관해 기술한다.

2.1 공간 데이터 웨어하우스와 공간 OLAP 연산

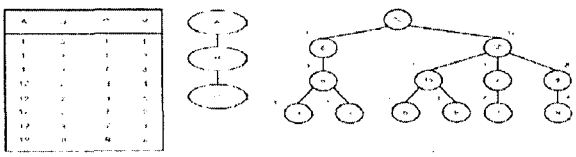
공간 데이터 웨어하우스는 공간과 비공간 데이터를 주제별로 통합하여 지리정보에 대한 의사결정을 효율적으로 지원하기 위한 시스템이다. 공간 데이터 웨어하우스는 차원(dimension) 테이블과 사실 테이블(fact column)이 공간 속성 데이터와 비공간 속성 데이터 모두를 포함하고 있기 때문에 공간 데이터에 대한 다차원 분석을 지원할 수 있

1) 본 연구는 대학 IT 연구센터 육성 · 지원사업의 연구결과로 수행되었음

다[2]. 또한 공간 데이터에 대한 효율적인 공간 OLAP 연산 지원을 위한 모델링 스키마로 비공간 데이터 웨어하우스에서 사용되는 스타/스노우플레이크 스키마(Star/Snowflake Schema)[1]가 사용된다. 공간 OLAP 연산은 보통 서로 다른 차원의 서로 다른 레벨에서 커다란 공간 객체를 요약한다. 그러므로 공간 사실 컬럼을 계산 하는 비용이 크기 때문에 공간 사실 컬럼에 선택적 실체화 기법을 적용한다. 또한 공간 OLAP 연산을 지원하기 위해 R-Tree 색인 구조를 유지한다. R-Tree는 점이나 선, 면 등의 다양한 공간 데이터를 처리하기 위해 MBR(Minimum Bounding Rectangle)을 이용해 데이터를 표현한다. R-Tree에서는 노드들의 MBR이 겹칠 수 있으므로 하나의 검색을 처리하기 위해 여러 노드들을 방문하게 될 수도 있다.

2.2 OLAP 연산 지원을 위한 계층구조 색인

OLAP은 의사 결정자가 다차원적인 관점에서 데이터를 분석할 수 있도록 해준다. 의사 결정 질의는 다양하고, 동적인 요청에 대해 빠른 응답시간을 요구한다. 그러나 일반적으로 대부분의 OLAP 질의는 특정 레코드를 다루는 것 보다는 전반적인 동향을 분석하기 위한 것으로 하나 이상의 집계(aggregate) 함수와 group-by 연산자가 포함되기 때문에 사용자의 기대치보다 응답 시간이 매우 오래 걸린다.



[그림 1] 큐브 트리 템플릿과 실체화

질의응답 시간을 향상시키기 위해 다차원데이터를 릴레이션 형태로 저장한 후 다차원 트리 형태의 인덱스 구조를 통해 각 다차원 데이터를 효율적으로 접근할 수 있도록 한다. 다차원 데이터는 일반적으로 성기기 때문에(sparse) 트리 구조를 사용하면 공간을 효율적으로 사용할 수 있다[6]. 트리 기반의 범위 집계 질의에서는 트리의 중간 노드에 집계 값을 저장하여 트리의 단말 노드를 접근하지 않고도 범위 내의 데이터에 대한 집계 값을 구할 수 있다. 각 노드는 한 차원에 대한 인덱스를 나타내고, 부모 노드는 자식 노드들의 합계를 저장하고 있다.

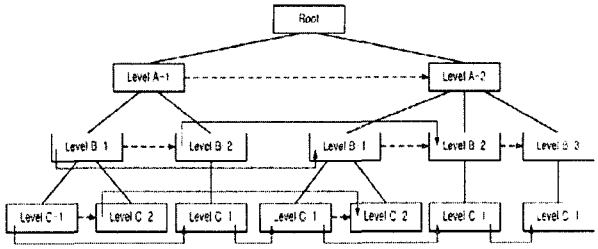
큐브 트리[6]는 차원들의 순서를 나타내는 템플릿에 의해 구체화 되어진다. [그림 1]의 왼쪽 편은 테이블 R과 A, B, C 차원 순서로 인덱스된 템플릿을 나타내며, 오른쪽은 템플릿에 의해 실체화된 큐브 트리의 예를 나타낸다. 각 리프노드 안의 수는 A, B, C에 해당하는 사실 컬럼 V의 합을 나타낸다. 그 위의 부모 노드 안의 수는 AB 조합에 의한 사실 컬럼 V의 합을 나타낸다. 루트 레벨에서는 전체 릴레이션에 사실 컬럼 V의 합을 나타낸다.

3. 분포 지역 질의 처리를 위한 큐브 트리 확장 기법

이 장에서는 제안 기법인 공간 연산 지원을 위한 확장된 큐브 트리를 설명한다. 관련연구에서 살펴본 큐브 트리는 차원 테이블에서 각 계층의 합(sum)을 계산하는 것에 초점이 맞추어져 있었다. 본 논문에서는 분포지역을 요청하는 범위 질의를 처리하기 위한 구조로 트리를 확장하고, 비공간 속성 질의가 공간 속성의 결과 값으로 변환되는 과정을 기술한다.

3.1 공간 연산 지원을 위해 확장된 큐브 트리

제안하는 확장된 큐브 트리의 구조는 [그림 2]와 같으며, 큐브를 구성하는 차원 테이블의 계층 구조 레벨에 따라서 트리의 깊이가 결정된다. 즉 각 레벨은 차원 테이블에서의 계층 구조를 표현하며, [그림 2]에서와 같이 각 레벨에서 범위 탐색을 지원하기 위해 각 형제 노드들(Level A-1의 Level B-1, Level B-2) 사이의 탐색을 지원하고, 각 레벨의 유사한 속성 도메인을 가진 탐색을 지원하기 위해 다른 부모의 자식 노드들(Level A-1의 Level B-1과 Level A-2의 Level B-1) 사이의 탐색도 지원한다. 또한 트리의 탐색 결과인 공간 객체에 대한 포인터들을 유지하며, 근접한 MBR을 가진 지역들에 대해서는 병합된 공간 객체로 관리함으로써 포인터의 수를 줄일 수 있고, 논리적인 분석 단위로 유지할 수 있다.



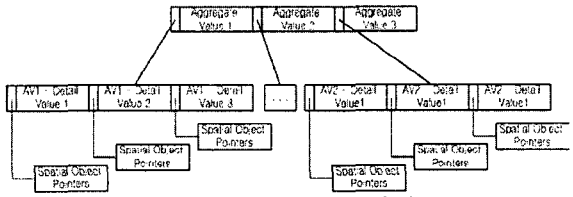
[그림 2] 확장된 계층 구조 트리의 구조

3.2 자료구조

각 노드는 [그림 3]과 같은 자료 구조를 유지한다. *Hierarchical Fact Column*은 차원 테이블의 계층에 의해서 결정되는 사실 컬럼(Fact Column)의 값이고, *Child Pointer*는 자식노드에 대한 포인터이다. *Next Sibling Pointer*는 다음 형제노드를 가리키며 범위 탐색을 지원한다. *Next Sibling Pointer*를 유지함으로써 부모노드에서 다시 탐색되어져야 하는 오버헤드를 줄일 수 있다. *Next Level Pointer*는 다음 서브 트리의 유사한 속성 도메인을 가진 노드를 가리키며 유사 속성 노드 탐색을 지원한다. *Next Level Pointer*를 유지함으로써 서로 다른 부모 노드를 가진 서브 트리의 탐색을 피할 수 있다.

[그림 3] 노드의 자료구조

*Hierarchical Fact Column*은 사실 컬럼을 계층화시킨 것이며, [그림 4]와 같은 자료구조를 가지고 있다. 각 *Aggregate Value 1, Aggregate Value 2, Aggregate Value 3*는 사실 컬럼의 계층화된 최상위 값이며, 그 하위의 *AV1-Detail Value 1, AV1-Detail Value 2, AV1-Detail Value 3*은 각 *Aggregate Value*에 대한 상세 값을 가진다. *Spatial Object Pointers*는 그 계층화에 해당하는 공간 객체 포인터들을 관리하고 있다. 이를 유지함으로써 의사 결정자가 분포지역의 결과에 대한 원인을 분석하고자 할 때, 분석하고자 하는 공간 데이터에 접근해 의사 결정에 필요한 정보와 특징들을 활용해 좀 더 정확한 의사 결정을 내릴 수 있게 한다. 또한 분포 질의의 특성상 MBR을 사용해 공간 데이터 객체들을 관리하기가 부적합하기 때문에 본 논문에서의 공간 객체들은 R-Tree 색인 기법으로 유지되지 않고 포인터들의 연결 리스트 구조로 유지된다.



[그림 4] Hierarchical Fact Column의 자료 구조

3.3 질의 처리

분포 지역을 요청하는 질의를 처리하는 과정의 알고리즘은 [그림 5]와 같다.

Step 1. 큐브 트리에서 질의를 처리할 수 있는 레벨 찾기
*Step 2. Hierarchical Fact Column*에서 사용자가 요청한 검색 조건과 맞을 때 까지 트리를 탐색
Step 3. 탐색된 노드에 존재하는 공간 데이터 분석 및 처리
Step 4. 범위 검색 질의와 속성 탐색 질의의 형태에 따라 *Next Level Pointer* 또는 *Next Sibling Pointer*를 따라 범위 탐색을 수행
Step 5. 질의 범위를 모두 탐색할 때 까지 *Step2, Step3* 과정을 반복

[그림 5] 질의 처리 알고리즘

[그림 5]에서와 같이 질의 처리는 각 분석 질의의 형태에 적합한 포인터를 따라 질의 범위에 해당하는 노드 탐색을 하며 결과 값을 얻을 수 있다. [그림 6]과 같이 시간 차원 계층을 가지는 차원 테이블이 존재한다고 가정하고 이 테이블을 계층 구조 트리로 구성을 하면 [그림 7]과 같은 형태의 트리가 구축된다. 범위 검색 질의와 동일 속성 도메인에 대한 검색 질의가 의사 결정자로부터 가장 빈번하며, 이 두 가지 질의 예제를 통해서 질의 처리 과정을 설명하겠다.

Year	Month	Day	Sales Quantity
2002	1	1	29
2002	1	...	...
2002	1	31	53
2002	...	...	...
2002	...	...	..
2003	12	1	11
2003	12	...	...
2003	12	31	70

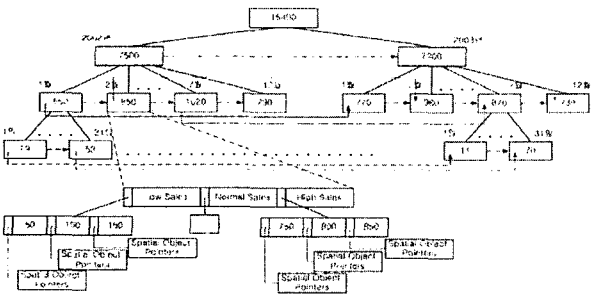
[그림 6] 시간 계층을 가진 차원 테이블

- 범위 검색 질의 처리 과정

  범위 검색 질의는 큐브 트리에서 질의를 처리할 수 있는 레벨을 찾고 질의 검색 조건과 맞으면 *Next Sibling Pointer*를 따라 질의 범위에 해당하는 노드를 탐색하며 결과 값을 얻는다. "2002년 2월부터 7월까지 판매량이 가장 높은 지역을 찾으시오"와 같은 예제를 통해 질의 처리 과정을 살펴보자. 2002년 2월부터 7월까지의 범위 검색을 수행하기 위해 Month 레벨로 트리를 탐색하게 되고 각 노드의 *Hierarchical Fact Column*을 통해 판매량이 높은 지역들을 탐색한 후 *Next Sibling Pointer*를 따라가며 반복 수행한다. 노드의 순차 탐색을 지원하기 때문에 효율적인 검색을 지원한다. 이런 과정을 통해 판매 수량이 높은 지역을 알 수 있고, 이 노드가 포함하고 있는 공간 객체들을 분석해 판매 수량이 높은 이유를 분석할 수 있으며 다른 지역에 대해서도 계획을 수립할 수 있다.

- 동일 속성 도메인에 대한 검색 질의 처리 과정

  동일 속성 도메인에 대한 검색 질의는 큐브 트리에서 질의를 처리할 수 있는 레벨을 찾고 질의 검색 조건과 맞으면 *Next Level Pointer*를 따라 질의 범위에 해당하는 노드를 탐색하며 결과 값을 얻는다. "2002년부터 2003년까지 2월 판매량이 900건을 넘는 지역을 찾으시오"와 같은 예제를 통해 질의 처리 과정을 설펴보자. 2002년부터 2003년까지 2월 판매량이므로 2002년의 2월 노드까지 이동하고 *Hierarchical Fact Column*을 통해 판매량이 900건이 넘는 지역들을 탐색한 후 *Next Level Pointer*를 따라서 2003년 2월 노드를 탐색할 수 있다. 이런 탐색을 통해 판매 수량이 900건이 넘는 노드를 선별할 수 있으며, 선별된 노드들의 공간 객체 포인터를 사용해 사용자가 원하는 분석을 수행할 수 있다.



[그림 7] 확장된 계층 구조 트리의 검색 예제

공간 데이터 웨어하우스 시스템에 제안 색인 기법을 구축함으로써 비공간 속성 질의를 통해 공간 객체를 결과로 요청하는 형태의 질의를 지원할 수 있게 되며 사실 컬럼을 계층화시킴으로서 사용자에게 좀 더 다각적인 분석을 지원할 수 있다.

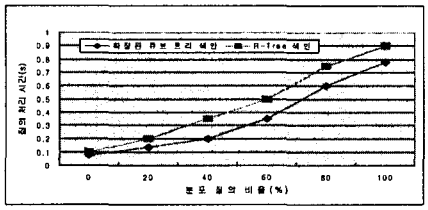
4. 성능 평가

본 장에서는 본 연구실에서 개발한 공간 데이터베이스 시스템인 GMS에서 R-Tree 색인 기법과 본 논문에서 제안하는 확장된 큐브 트리 색인 기법에 대한 성능 평가를 하였다. 테스트 환경은 [표 1]과 같으며, 일반 공간 검색 질의에 분포 질의의 비율을 높이면서 실시했다.

[표 1] 시간 계층을 가진 차원 테이블

컴퓨터	IBM PC Compatible
CPU	Intel Pentium 4, 2.4GHz
메모리	768MB
디스크	7200RPM, 80GB
운영체제	MS Windows 2000
개발환경	MS Visual C++ 6.0

성능 평가 결과를 통해 분포 지역 질의인 범위 검색 질의와 동일 속성 도메인에 대한 검색 질의의 비율이 늘어날수록 제안 기법을 사용했을 때 질의 처리 성능이 약 15%정도 향상됨을 알 수 있다.



[그림 8] 확장된 큐브 트리 색인과 R-Tree 색인의 성능 비교

5. 결론 및 향후 연구

본 논문에서는 공간 데이터 웨어하우스에서 비공간 속성 질의를 공간 속성 결과로 변환하는 분포 지역 요청 질의를 처리하기 위해 확장된 큐브 트리를 제안하였다. 제안 기법은 분석하고자 하는 사실 테이블의 비공간 속성을 큐브 트리의 키로 사용하고, 단말 노드에 동일한 비공간 속성 값을 가진 공간 데이터에 대한 포인터들을 관리함으로서 비공간 속성 질의를 공간 속성 결과로 변환한다. 제안 기법을 공간 데이터 웨어하우스 시스템에 구축함으로써 비공간 속성 질의를 통해 공간 객체를 결과로 요청하는 형태의 질의를 노드의 탐색 비용을 줄이면서 지원할 수 있게 된다. 또한 큐브 트리를 기반으로 하기 때문에 OLAP 연산에 필수적인 계층 구조를 지원할 수 있으며 사실 컬럼을 계층화시킴으로서 사용자에게 좀 더 다각적인 분석을 지원할 수 있다.

향후 연구로는 본 논문이 제안한 기법에서의 공간 포인터의 효율적인 관리와 다차원 계층 구조 트리를 이용한 사용자의 다차원 분석 질의 처리 수행 기법에 대해서 연구하겠다.

참고 문헌

[1] S. Chaudhuri and U. Dayal, An Overview of Data Warehousing and OLAP Technology, ACM SIGMOD Record 26(1), pp. 65-74, 1997.

[2] J Han, N Stefanovic, and K Koperski, Selective Materialization : An Efficient Method for Spatial Data Cubc Construction, PAKDD'98, pp. 144-158, 1998.

[3] D. Papadias, P.Kalnis, J. Zhang, and Y. Tao, Efficient OLAP operations in spatial data warehouses, Lecture Notes in Computer Science, 2001.

[4] F. Rao, L. Zhang, X. L. Yu, Y. Li, and Y. Chen, Spatial hierarchy and olap-favored search in spatial data warehouse, DOLAP'03, 2003.

[5] A. Guttman, R-Trees : A Dynamic Index Structure for Spatial Searching, ACM SIGMOD, pp. 47-57, 1984.

[6] T. johnson and D. Shasha, Some approaches to index design for cube forest, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 20(1), pp. 27-35, 1997.