韓國 OA學會 論文誌
第3卷 第4號, 1998. 12.

98-3-4-3-8

효율적인 공간 데이타 웨어하우스에 관한 연구

이 기 영*

A Study on the Effective Spatial Data Warehouse

Ki-Young Lee*

요 약

최근 의사결정 지원 시스템의 필요성과 중요성이 크게 대두되어 활발히 연구가 진행되고 있는 공간 데이타 웨어하우스는 요약·분석 작업을 통해 의사 결정에 필요한 정보를 사용자에게 효율적으로 제공하기 위하여 과거로부터 누적된 이질적이고 다양한 대용량의 공간 데이타로 구성된다. 본 논문에서는 공간 데이타 웨어하우스 모델링을 위해서 가상적인 시나리오를 토대로 효과적인 공간 다차원 모델을 고찰한다. 또한 사용자가 요구하는 기본적이고 핵심적인 분석 기능들을 모두 표현할 수 있는 공간 다차원 분석 질의의 여러 형태에 대하여 살펴본다.

Abstract

Spatial data warehouse, whose importance is being increased, is composed of huge amounts of historical spatial data for organizational decision making and it also allows users to obtain useful geospatial information through analyzing and summmarizing spatial data. In this paper, we survey effective spatial multidimensional model which is based on virtual scenario for spatial data warehouse modelling. Therefore, we describe spatial multidimensional analytical query which provide multiple analytical functions according to user's requests.

* 서울보건대학 사무자동화과 조교수
논문접수 : 98.11.14. 심사완료 : 98.12.18.

I. 서 론

최근 들어 의사결정 지원 시스템(DSS: Decision Support System)의 필요성이 증가함에 따라 본격적으로 관심을 모으기 시작한 분야가 데이타 웨어하우스(Data Warehouse)이다 [1, 2,]. 주로 기업, 관공서 등과 같은 방대한 양의 정보를 다루는 기관에서 비공간 데이타를 통합된 체계로 관리함으로서 의사결정 지원을 효율적으로 지원할 수 있도록 데이타 웨어하우스에 대한 연구가 활발히 진행되고 있다 [11, 13]. 그러나 데이타 웨어하우스에 구성된 데이타들의 유형을 살펴볼 때에 데이타의 80퍼센트 이상이 소비자 주소, ZIP 코드, 상점 위치, 지역 영역 등과 같은 공간적인 요소를 내포하고 있음에도 불구하고 데이타 웨어하우스에 이러한 공간적인 요소들을 표현, 통합하여 관리할 수 있는 공간(Spatial) 데이타 웨어하우스에 관한 연구는 아직 미흡한 실정에 있다 [3, 5, 7, 10].

대용량의 공간 데이타를 포함하는 공간 데이타 웨어하우스의 구축을 위해 공간 데이타 웨어하우스에서는 개념 설계보다 하위 단계 모델 방법으로 차원 모델(Dimensional Model)을 사용한다 [8, 9]. 그러므로 공간 데이타 웨어하우스의 주요 업무 처리가 공간적 분석 작업에 해당하므로 분석의 대상이 되는 응용분야의 개념들에 기초한 다차원(Multidimensional)적인 관점을 제공하는 것이 중요하다.

의사결정 지원 시스템으로서 공간 데이타 웨어하우스를 생각할 때 사용자에게 공간 데이타에 대한 다차원 관점을 제공하는 것은 효율적이고 바른 분석을 의한 핵심 기술이 된다. 아직까지는 다차원 관점의 데이타와 분석 질의에 대한 모델이 정립되지 않아 기존의 시스템들은 다차원 관점 질의 인터페이스를 나름대로의 형식으로 제공하고 있는 실정이다. 그러므로 사용자가 요구하는 기본적이고 핵심적인 분석 기능들을 모두 표현해 낼 수 있는 공간 다차원 분석 질의 모델이 절실히 요구된다 [4, 14].

본 논문에서는 공간 데이타 웨어하우스 모델링을 위해서 가상적인 시나리오를 토대로 효과적인 공간 다차원 모델을 제시하였다. 또한 사용자가 요구하는 기본적이고 핵심적인 분석 기능들을 모두 표현할수 있는 공간 다차원 분석 질의의 여러 유형에 대하여 정의하였다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 관련 연구로서 데이타 웨어하우스를 위하여 기존에 제안된 다차원 모델의 방법들과 다차원 분석 질의에 대해 살펴보고, 제 3 장에서는 공간 데이타 웨어하우스의 효율적인 공간 다차원 모델링을 위해 가상적 시나리오와 이 시나리오를 토대로 공간 다차원 모델을 제시한다. 제 4 장에서는 공간 데이타 웨어하우스의 공간 다차원 분석 질의 유형들을 정의하고 예제를 통해서 살펴본다. 마지막으로 제 5장에서는 결론과 앞으로의 연구 과제에 대해 언급한다.

II. 관련연구

본 장에서는 공간 데이타 웨어하우스는 기존 데이타 웨어하우스가 가지는 특징 [6] 외에 또 다른 특징을 포함하여야 하므로 효율적인 공간 다차원 모델을 제시함에 있어 기존 다차원 모델 방법과 다차원 분석 질의에 대하여 살펴본다.

2.1 기존 다차원 모델

데이타 웨어하우스를 위해 기존에 제안된 다차원 모델의 방법들은 정규화 정도에 따라 스타 스키마, 스노우플레이크 스키마, 사실 성좌 스키마, 혼합형 스키마, 갤럭시 스키마로 분류할 수 있으며 각 테이블에 대한 정규화 측면에서 살펴보면 다음과 같다.

(1) 스타 스키마(Star Schema)

스타 스키마는 하나의 사실 테이블과 하나 이상의 차원 테이블들로 구성된 트리 구조의 스키마이다 [12]. 모든 테이블에 대한 정규화 과정이 없기 때문에 데이타들의 중복 저장으로 인한 방대한 저장 공간이 필요할 뿐만 아니라, 차원 테이블의 크기가 큰 경우에는 비정규화 과정으로 인한 검색 시간의 비효율성과 죠인 연산이 복잡하

다.

그러나 각 테이블에 대한 비정규화로 데이타 모델의 구조를 간단하게 만들게 됨으로서 사용자에게 직관적이고 이해하지 쉬운 다차원 관점을 제공할 뿐만 아니라 테이블에 대한 탐색이 쉬워지고 테이블 관리 및 메타 데이타 관리가 수월하다.

(2) 스노우플레이크 스키마
(Snowflake Schema)

스노우플레이크 스키마는 스타 스키마에 존재하는 각각의 사실 테이블과 차원 테이블을 모두 정규화시킨 스키마이다. 모든 테이블에 대하여 정규화를 시켰기 때문에 테이블의 크기가 작고 저장 공간이 스타 스키마보다 적게 필요하다.

(3) 사실 성좌 스키마
(Fact Constellation Schema)

사실 성좌 스키마는 스타 스키마에 존재하는 사실 테이블을 집계 별로 정규화시킨 스키마이다. 사실 테이블만을 정규화를 시켰기 때문에 테이블의 크기가 스타 스키마보다는 작고 저장 공간 요구량은 스타 스키마보다는 적으나 스노우플레이크 스키마보다는 크다. 사실 테이블에 대한 질의 성능면에서는 우수성을 갖는다.

(4) 혼합형 스키마(Mixed Schema)

혼합형 스키마는 스타 스키마, 스노우플레이크 스키마를 절충한 스키마로서, 다만 차원 테이블이 아주 큰 경우에만 그 테이블을 정규화시킨 스키마이다. 그러므로 스타 스키마, 스노우플레이크 스키마에서 비정규화, 정규화로 인한 장점과 단점들을 절충시킨 스키마이다.

(5) 갤럭시 스키마(Galaxy Schema)

갤럭시 스키마는 스타 스키마에서는 오직 하나의 사실 테이블만을 갖는데 반하여 이 스키마에서는 여러개의 사실 테이블들을 갖을 수 있는 스키마로서 비정규화 시킨 스키마이다.

본 논문에서는 공간 데이타 웨어하우스의 효율적인 공간 다차원 모델을 제시함에 있어 공간 데이타의 기억 공간을 최소화하기 위해 기존의 다차원 모델을 확장하여 적용하였다.

2.2 다차원 분석 질의

방대한 양의 데이타를 읽거나 집계하여 데이타간의 복잡한 관계나 패턴, 경향 등을 분석하는 다차원 분석 질의는 다음과 같은 집계(Aggregation), Drill-down, Drill-up, Drill-across, 슬라이싱(Slicing), 다이싱(Dicing), 피보팅(Pivoting), 다차원 관점 제공 등의 기본적인 기능을 제공해야 하며, 또한 의사 결정을 지원하기 위한 특수 함수 등을 지원하여야 한다 [14].

공간 데이타 웨어하우스에서의 질의는 기존의 처리 방법만으로는 효율적인 처리가 불가능하며 공간 데이타 웨어하우스의 특성에 부합하는 새로운 공간 질의 처리 시스템의 모델이 정립되어야 한다. 또한, 공간 분석 질의는 SQL과는 달리 복잡하고 다양한 연산을 요구하므로 이에 대한 형식화된 모델의 정립이 요구된다.

III. 공간 다차원 모델

본 장에서는 효율적인 공간 데이타 웨어하우스 구축 및 설계를 위하여 스키마에서 발생하는 저장 공간의 낭비 문제에 따른 시스템 효율의 저하를 원활히 해결하기 위해 기존의 다차원 모델을 확장한 공간 다차원 모델을 가상적인 시나리오 측면에서 고찰한다.

3.1 공간 다차원 모델링을 위한 시나리오

〈시나리오〉 서울시 각 구청에서는 각각의 지리 정보 시스템을 구축하여 활용하고 있으며, 각 구청에서 사용하는 이러한 시스템들은 공간 데이타와 공간 데이타와 밀접한 관계가 있는 비공간 데이타(건축, 인구, 수송, 관광사업, 그외 지리적 환경과 밀접한 다른 경제 활동들)로 데이타베이스를 구축하고 있다. 또한 구청의 각 부서에서는 부서의 업무에 적합한 데이타(공간, 비공간)를 갖고 운영한다. 그러나 최근 들어 의사결정 지원 시스템의 필요성이 대두되어 우선 시험적으로 "건축과"에서만 공간 데이타 웨어하우스를 구축하고자 한다.

위와 같은 시나리오는 공간 데이타 웨어하우스를 구축

하기 위해 적용될 수 있는 환경을 제공하고 있다. 그러므로 본 논문에서는 이 시나리오를 토대로 하여 공간 다차원 모델링을 언급하였다.

3.2 시나리오를 위한 공간 다차원 모델링

3.2.1 의사 결정 처리 사항과 사실 테이블의 식별

건축과의 관리자는 담당하고 있는 구의 모든 행정구역에서 특정 시점에서 준공된 건축물들을 분석하려고 한다. 이러한 처리 사항을 수행하려면 관리자가 중요하게 여기는 공간, 비공간 데이타는 건축물 수와 그 건축물의 공간 위치이므로 사실 테이블은 건축물 수와 공간 위치 테이블이 된다. 그러므로 사실 테이블의 사실은 "건축물 수"와 "공간 위치"가 집계될 것이다. "공간 위치"는 효율적인 공간 질의 처리를 위해 건축물들의 위치를 가르키고 있는 공간 포인터로 표현한다. 또한 사실 테이블은 사실 테이블에 관련된 모든 차원 테이블의 결합 키로 구성된다.

3.2.2 차원 테이블 결정

사실 테이블은 시간, 건축물, 행정구역에 따라 분석되어야 하므로, 시간, 건축물, 행정구역 차원의 3개의 차원 테이블로 구성된다. 그러나 공간 데이타의 특성으로 인하여 공간 데이타는 방대한 기억 공간을 차지하므로 행정구역 차원 테이블은 행정구역별("구", "동", "가" 지역)로 정규화시킨다.

3.2.3 차원 테이블의 속성 결정

시나리오에 대한 공간 다차원 모델의 구성 문제를 단순화하기 위해 차원 테이블의 상세한 속성 설계는 생략하기로 한다. "시간" 차원 테이블은 시간 키와 함께 일, 주, 월, 회계분기, 연도 등이 포함된다. 그러나 공간 데이타 웨어하우스를 구축 시에 조직의 활동 상황에 따라 시간 개념의 효율적인 모델링이 필요하다. "건축물" 차원 테이블은 건축물 키와 함께 건축물의 위치를 표현하기 위한 공간 데이타(1차원-점)와 이 공간 데이타의 위상을 나타내는 속성 데이타와 비공간 데이타(건축물설명, 소유주, 평수, 공시가 등)로 구성된다. 행정구역 차원 테이블은 행정구역별("구", "동", "가" 지역)로 정규화시켰기 때문에 "구" 행정구역, "동" 행정구역, "가" 행정구역 차원 테이블이 있다.

"구" 행정구역 차원 테이블은 "구" geocode 키와 함께 이 건축물이 어느 동에 속하는지를 나타내기 위한 "동" geocode 키를 포함한다. 또한 "구"의 공간 위치를 표현하기 위한 공간 데이타(2차원-다각형)와 이 공간 데이타의 위상을 나타내는 속성 데이타와 비공간 데이타(구설명, 면적 등)로 구성된다. "동" 행정구역 차원 테이블은 "동" geocode 키와 함께 이 건축물이 어느 가에 속하는지를 나타내기 위한 "가" geocode 키를 포함한다. 또한 "동"의 위치를 표현하기 위한 공간 데이타(2차원-다각형)와 이 공간 데이타의 위상을 나타내는 속성 데이타와 비공간 데이타(동설명, 면적 등)로 구성된다.

"가" 행정구역 차원 테이블은 "가" geocode 키와 함께 "가" 위치를 표현하기 위한 공간 데이터(2차원-다각형)와 이 공간 데이타의 위상을 나타내는 속성 데이타와 비공간 데이타(가설명, 면적 등)로 구성된다. 조직의 응용에 따라 공간 데이타의 효과적인 표현 기법이 필요하다.

3.2.4 효율적인 공간 다차원 모델

2. 1 절에서 언급한 바와 같이 기존에 제안된 다차원 모델의 방법들은 정규화 정도에 따라 여러 가지 분류가 있으나 시나리오에 적합한 효율적인 공간 다차원 모델링을 위해서는 스키마에서 발생하는 저장 공간의 낭비, 죠인 문제에 따른 시스템 효율의 저하를 원활히 해결해야 하고, 효율적인 질의 처리를 보장할 수 있어야 한다. 그러므로, 본 논문에서는 위의 절차에 따른 가상적 시나리오를 위한 공간 다차원 모델의 스키마는 그림 1.과 같다.

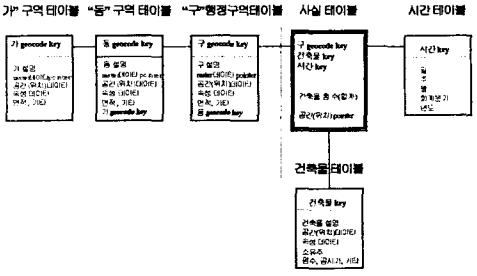

그림 1. 시나리오를 위한 공간 다차원 모델

Ⅳ. 공간 다차원 분석 질의 유형

3. 2절에서 언급한 시나리오에 대한 공간 다차원 모델 스키마를 중심으로 공간 다차원 분석 질의 유형들을 개념적으로 고찰하기 위해 질의 유형의 형태와 질의 결과 유형을 표기식으로 정의하면 다음과 같다.

(정의) 공간 다차원 분석 질의 유형

{ ⟨T⟩*, ⟨S , NS⟩*, ⟨C⟩* }Q_t ⇒ {⟨S , NS⟩* , ⟨T⟩*}R_t

단, 위의 정의에서 표현된 T는 공간 다차원 분석 질의와 그 질의 결과 내용중에 나타나는 시간적인 도메인 요소, S는 공간적인 도메인 요소, NS는 비공간적인 도메인 요소를 의미하고, C는 질의 제약 조건적인 요소를 의미한다. ⟨ ⟩기호는 각 괄호안의 도메인 요소들이 조합적으로 나타날 수 있음을 의미하며 *는 분석 질의와 그 질의 결과 내용중에 나타나는 임의의 도메인 요소들이 반복적으로 나타날 수 있음을 의미한다. 또한 위 정의에서 ,는 임의의 여러 도메인 요소들이 and 또는 or로 나타날 수 있음을 의미하고 Q_t는 질의할 수 있는 모든 공간 다차원 분석 질의중 하나의 질의 유형을 나타내며 R_t는 공간 다차원 분석 질의(Q_t)에 대한 질의 결과의 한 유형의 형태를 의미한다

위에서 정의한 공간 다차원 분석 질의 유형을 3. 1절에서 언급한 시나리오 상황에 적용한 질의 예는 다음과 같으며 모든 공간 다차원 분석 질의 유형과 그 질의 결과에 적용할 수 있다

(질의 예1) 특정 구청 건축과의 관리자는 담당하고 있는 구에서 어느 시간에 준공된 건축물의 수 또는 건축물 위치를 분석하고자 한다. 그래서 그 관리자는 "이 지역(광진구)"에서 "최근 3년간(95-97)"에 준공된 "60평 이상의 편의점 위치" ?를 분석하고 싶어서 이와 같은 질의를 하였다.

위에서 언급한 정의에 적용하기 위해 위의 질의를 분석하면 "이 지역"은 공간적 도메인 요소이고,"최근 3년간"은 시간적 도메인 요소, "60평 이상"은 질의 제약 조건적 요소이다. 또한 질의 결과의 "편의점 위치"는 공간적 도메인 요소로 이루어진다. 그러므로 공간 다차원 분석 질의 유형은 { T , S , C }Q_t로 정의할 수 있고 그 질의 결과 유형은 { S }R_t로 정의할 수 있다.

(질의 예2) "이 지역(광진구)"에서 "최근 3년간(95-97)"에 준공된 "60평 이상의 편의점과 노래방의 위치 및 소유주와 공시가" ?를 분석하고 싶어서 위와 같은 질의를 하였다.

위의 질의 예에서는 "편의점"과 "노래방"의 위치는 공간적 도메인 요소들이고 "소유주"와 "공시가"는 비공간적 도메인 요소들이다. 그러므로 공간 다차원 분석 질의 유형은 { T , S , C }Q_t로 정의할 수 있고 그 질의 결과 유형은 { ⟨S , NS⟩* }R_t로 정의할 수 있다.

V. 결론 및 향후 연구 과제

데이타 웨어하우스 환경에서의 다차원 모델링 기법을 분석하여 볼 때 적합한 공간 다차원 모델은 단순성에 중점을 두면 스타 스키마가 적합하다고 판단된다. 그러나 정규화 기법은 테이블이 매우 큰 경우에 저장 공간 사용과 처리 속도상의 장점이 있으므로 공간 차원 테이블의 특징에 따라 부분적으로 정규화하는 절충적 기법이 가장 바람직하다.

본 논문에서는 공간 데이타를 위한 공간 데이타 웨어하우스의 설계와 구축을 위해 가상적인 시나리오를 만들어 이 시나리오를 토대로 하여 공간 데이타의 저장 공간과 질의 성능면을 향상시킨 효율적인 공간 다차원 모델을 제시하였고 공간 데이타 웨어하우스에 대한 공간 다차원 질의 유형의 형태에 대하여 고찰하였다.

향후 연구 방향으로는 사용자에게 보여지는 공간 다차원 관점에 대한 최적의 공간 다차원 모델과 사용자가 요구하는 기본적이고 핵심적인 분석 기능들을 모두 표현해 낼 수 있는 공간 다차원 분석 질의 모델에 대한 연구가 필요하다.

참고문헌

[1] Barguin, R.C. and Edelstein, H.A., "Planning and Designing the Data Warehouse", Prentice Hall, Inc., 1997.

[2] Butler Group, "Data Warehouse Issues", White Paper of Burler Group, 1997.

available via http://www.butlergroup.co.uk.

[3] ESRI, "Spatial Data Warehouse", White Paper of ESRI, 1997, available via http://www.esri.com.

[4] Gupta, A., Harinarayan, V., and Quass, D., "Aggregate-Query Processing in Data Warehousing Environments", Proceedings of the 21st VLDB Conference Zurich, Swizerland, pp. 358-369, 1995.

[5] Han, J., " Spatial Data Mining and Spatial Data Warehousing", SSD'97 Conference Tutorial, 1997, available via http://db.cs.sfu.ca.

[6] Inmon, W.H., "Building the Data Warehouse", 2nd Edition, John Wiley & Sons,Inc., 1996.

[7] Keighan, E., "The Mercator Geospatial Data Warehousing Architecture", Proceedings of The Twelfth Annual Symposium on Geographic Information Systems, Toronto, pp.184-187, May 1998.

[8] Kelly, T.J., "Dimensional Data Modeling", White Paper of Sybase,1997, available via http://www.sybase.com.

[9] Labio, W., Wiener, J., Gupta, H., Garcia-Molina, H., and Widom, J., " The WHIPS Prototype for Data Warehouse Creation and Maintenance", Stanford Univ., 1997, available via http://www-db.stanford.edu.

[10] Lafond, P., "Designing and Building the Distributed Geospatial Data Warehousing Architecture", Proceedings of The Twelfth Annual Symposium on Geographic Information Systems, Toronto, pp.188-190, May 1998.

[11] Nebraska Univ. White Paper, " Nulook Data Warehouse Project", 1997, available via http://www.nulook.uneb.edu.

[12] Red Brick System, " Star Schemas and STARjoin Technology", Red Brick Systems White Paper, 1996.

[13] Stanford Univ. White Paper, "The Data Warehousing Projects", 1996.

[14] 정유나, 장재영, 이상구, "데이타 웨어하우스 환경에서의 다차원 질의 모델",정보 과학회 논문집, 제 24권 제 2호, pp. 97 - 100, 11월 1997.

저 자 소 개

이기영

1984년 : 숭실대학교 전자계산학과 졸업(공학사)

1984년~1991년 : 한국해양연구소 전산연구실(연 구원)

1988년 : 건국대학교 컴퓨터공학과 (공학석사)

1991년 : 서울보건대학 전산정보처리과 전임강사

1992년~현재 : 서울보건대학 사무자동화과 조교수

1998년 : 건국대학교 컴퓨터공학과 박사과정수료

관심분야 : GIS, 데이터 마이닝, 데이터 웨어하우스