XML 소스 데이터로부터 스타 스키마를 생성하기 위한 XML2Star 알고리즘

최은하[0] 김진호
강원대학교 컴퓨터과학과
neva99@hanmir.com jhkim@kangwon.ac.kr

XML2Star Algorithm Creating Star Schema from Source Data in XML

Eun-Ha Choi[0] Jin-Ho Kim
Dept. of Computer Science, Kangwon National University

요 약

데이터 웨어하우스는 기업의 의사 결정을 지원하기 위해 기업의 운영 데이터베이스로부터 추출한 데이터의 집합으로써 OLAP 분석에 이용된다. OLAP은 데이터에 대한 다양한 분석을 위해 이들 데이터를 다차원 데이터 모델로 표현하고 이를 활용하여 복잡한 질의 처리 및 다차원 데이터 분석에 이용한다. 이러한 OLAP의 다차원 데이터를 관계형 데이터베이스에서 표현하기 위해 스타 스키마가 널리 사용된다. 지금까지의 데이터 웨어하우스는 일반적으로 ER 도형으로 설계된 소스 데이터로부터 스타 스키마를 설계하고 구축하였다. 하지만, 최근 인터넷의 급성장으로 인해 차세대 웹 문서의 표준인 XML을 통한 인터넷 상의 문서 전송 및 정보 교환이 활발해 지고 있으며, XML 문서에 대한 다차원적인 분석이 요구됨에 따라 데이터 웨어하우스는 XML 문서로부터의 스타 스키마 설계 및 저장이 필요하게 되었다. 따라서 본 논문에서는 XML DTD로부터 애트리뷰트 트리를 생성하여 스타 스키마를 설계하고, 이 DTD를 따르는 XML 문서에서 스타 스키마의 인스턴스를 추출하여 관계형 데이터베이스에 저장하기 위한 XML2Star 알고리즘을 개발하였다. 이것을 통해 기업 및 사용자는 OLAP에서 XML 기반의 스타 스키마를 이용한 다차원적인 분석이 가능하게 된다.

1. 서 론

실시간 분석 처리(On-Line Analytical Processing : OLAP)는 기업의 의사 결정 지원을 위하여 다차원적인 데이터 분석 및 복잡 질의 처리에 효율적이며, 데이터 웨어하우스는 OLAP 분석에 사용할 데이터를 저장하고 있다[1][2]. OLAP에서는 데이터를 분석에 적합하도록 다차원 데이터 형태로 모델화하게 되는데 이를 다차원 큐브라 하며, 관계형 데이터베이스에서 다차원 데이터 모델인 큐브를 표현하기 위해 등장한 데이터 구조가 스타 스키마이다. 스타 스키마는 하나의 사실 테이블과 여러 개의 차원 테이블로 구성되어 있고, 테이블간의 조인을 최소화하여 질의에 대한 응답시간을 줄일 수 있기 때문에 복잡한 질의에 적합한 데이터 구조이다[2][3].

이러한 데이터 웨어하우스의 설계에 관한 지금까지의 연구는 일반적으로 ER 도형으로 설계되어 있는 소스 데이터로부터 어떻게 효율적으로 스타 스키마를 설계하는가에 관한 것이었다. 하지만 최근 인터넷의 급성장으로 인한 인터넷에서의 문서 전송 및 정보 교환이 증가함에 따라 데이터 웨어하우스는 인터넷 상에서의 정보를 저장하고 구축하는 것이 필요하게 되었다. XML은 이러한 인터넷에서의 문서 전송 및 정보 교환을 용이하도록 하기 위해 W3C에서 제안한 차세대 웹 문서의 표준이다.[4] 이에 따라 기업 및 사용자는 의사 결정 지원을 위해 XML 문서에 대한 분석을 요구하게 되었으며, 데이터 웨어하우스는 XML 문서를 저장하여 웹 문서에 대한 다차원 분석을 제공하는 것이 필요하게 되었다.

본 논문에서는 XML 문서를 데이터 웨어하우스의 소스 데이터로 하고 이것으로부터 스타 스키마를 생성하기 위한 XML2Star 알고리즘을 개발하였다. 링크를 통해 연결되어 있는 XML 문서들, IDREF를 가지고 있는 XML 문서 등의 XML DTD로부터 애트리뷰트 트리를 생성하여 스타 스키마를 설계하고, 설계된 스타 스키마의 메타 정보를 이용하여 XML 문서를 관계형 데이터베이스에 저장하게 된다. 데이터 웨어하우스 설계자는 DTD를 통해서 XML 문서에 대한 스타 스키마의 구성 요소인 사실 및 측정값, 차원, 차원의 계층구조 등을 정의 할 수 있고, 이러한 DTD를 따르는 XML 문서로부터 스타 스키마의 인스턴스를 추출하여 저장할 수 있으며, 이것을 통해 기업 및 사용자는 OLAP 도구로부터 XML 문서에 대한 다차원적인 분석이 가능하게 된다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구와 본 논문의 연구 동기에 대해 서술하고, 3장에서는 XML2Star 알고리즘 중 XML DTD로부터 스타 스키마를 설계하는 과정을 설명한다. 4장에서는 설계된 스타 스키마를 이용하여 XML 문서를 관계형 데이터베이스에 저장하는 과정에 대해 설명한 후, 5장에서 결론을 맺는다.

* 이 논문은 첨단정보기술연구센터(AITrc)를 통하여 한국과학재단의 지원을 받았음.

2. 관련 연구

2.1 데이터 웨어하우스와 OLAP

데이터 웨어하우스는 의사 결정에 필요한 정보를 사용자에게 효율적으로 제공하기 위해 구축된 거대한 데이터 저장소이다. 이러한 데이터 웨어하우스에 근간을 두고 있는 OLAP은, 기업의 의사 결정 지원을 위하여 다차원적인 데이터 분석 및 복잡 질의 처리를 제공하는데 효율적이다. 데이터 웨어하우스 설계자는 여러 사이트로부터 통합된 소스 데이터와 사용자의 요구사항을 분석하여 데이터 웨어하우스를 설계하게 된다. M. Golfarelli와 S. Rizzi[5]는 데이터 웨어하우스 설계를 위한 방법론을 다음과 같은 단계로 설명하고 있다. 먼저, 운영 데이터베이스로부터 수집된 정보를 분석하고 사용자의 요구 사항을 명세하는 단계를 거친 후 개념적 설계를 통해 사실과 차원 구조를 생성하게 된다. 다음으로 개념적 설계 결과에 대한 작업량과 유효성을 판단한 후 테이블 구조로 실체화하기 위한 논리적 설계 단계를 거치고, 마지막으로 물리적 설계를 수행한다.

데이터 웨어하우스의 소스 데이터는 OLAP의 다차원 분석에 사용하기 위하여 분석에 적합한 다차원 데이터 모델 즉, 스타 스키마 형태로 변환되게 된다. 이에 따라 ER 도형, XML 등의 소스 데이터로부터 데이터 웨어하우스의 스타 스키마를 효율적으로 설계하기 위한 연구가 계속해서 진행되고 있다.

2.2 다차원 데이터 모델

다차원 데이터 모델 즉, 스타 스키마는 데이터 웨어하우스와 OLAP

의 출현으로 인해 요구되는 새로운 데이터 모델로써, 분석을 위한 다차원 데이터를 표현하기 위해 사실 테이블과 차원 테이블로 구성된다. 테이블간의 조인을 최소화 함으로써 질의에 대한 응답속도를 향상시킬 수 있다는 장점을 가지고 있기 때문에 OLAP의 복잡 질의에 적합하며, 데이터 웨어하우스 테이블 구조의 논리적 디자인에 사용되어 진다.[1][2][3]

L. Cabibbo 와 R. Torlone[3]은 ER 도형으로 설계되어 있는 소스 데이터로부터 다차원 그래프를 추출하고, 이 다차원 그래프로부터 다차원 데이터 모델을 설계하는 방법을 제시하고 있다. 또한 M. Golfarelli[6] 등은 ER 도형의 엔티티와 애트리뷰트를 노드로 가지는 애트리뷰트 트리를 생성하여 이것으로부터 다차원 데이터 모델을 설계하는 반자동적인 설계 방법을 제시하고 있다.

2.3 XML과 다차원 데이터 모델

최근 인터넷의 급속한 발전에 따라 인터넷 상에서의 문서 전송 및 정보 교환이 급증하고 있으며, 이러한 인터넷에서의 문서 전송 및 정보 교환을 용이하도록 하기 위해 차세대 웹 문서의 표준으로 제안된 것이 XML이다.

이에 따라 기업 및 사용자는 의사 결정 지원을 위해 XML 문서에 대한 분석을 요구하게 되었으며, 데이터 웨어하우스는 XML 문서를 저장하여 웹 문서에 대한 다차원 분석을 제공해야 할 필요가 생겨나게 되었다. 때문에 기존의 ER 도형으로 설계된 소스 데이터와 더불어 XML 소스 데이터로부터 다차원 데이터 모델을 설계하기 위한 연구가 활발히 진행되었다. M. R. Jensen[7][8]등은 XML 데이터를 UML 다이어그램으로 변환하여 다차원 데이터 모델인 스노우 플레이크 스키마를 생성하는 방법을 제안하였고, M. Golfarelli[9]등은 XML DTD로부터 애트리뷰트 트리를 생성하여 다차원 데이터 모델을 설계하는 방법을 제안하였다.

하지만 이들 논문에서는 DTD 1개에 대해서만 고려하였기 때문에, 실세계에서 여러 DTD를 따르는 XML 문서 사이에 링크가 있는 경우에는 부적합하다. 또한 엘리먼트 사이의 참조관계를 나타내는 IDREF 특성에 대한 언급이 없으며, 형제 엘리먼트 또는 하나의 엘리먼트 내의 애트리뷰트 사이의 계층구조를 정의하지 못하고 있다.

본 논문에서는 위의 문제점들을 고려하여 유효한 XML 문서로부터 스타 스키마를 생성하기 위한 XML2Star 알고리즘을 개발하였다. 이것을 통해 DTD로부터 스타 스키마를 설계하고, 설계된 스타 스키마의 메타 정보를 이용하여 DTD를 따르는 XML 문서를 관계형 데이터베이스에 저장한다.

3. 스타 스키마 설계

DTD를 따르는 유효한 XML 문서로부터 스타 스키마를 생성하기 위한 XML2Star 알고리즘은 DTD로부터 스타 스키마를 설계하는 단계와 설계된 스타 스키마의 메타 정보를 이용하여 XML 문서를 관계형 데이터베이스에 저장하는 단계로 이루어진다. 표 1은 XML2Star 알고리즘을 보여주고 있다.

표 1. XML2Star 알고리즘

1. DTD로부터 스타 스키마 설계
1-1. DTD 그래프 생성
1-2. 애트리뷰트 트리 생성
1-3. Attach/Graft/Prune
1-4. 스타 스키마 정의
2. 데이터 웨어하우스에 XML 문서 저장
2-1. 데이터 웨어하우스에 스타 스키마 정의
2-2. XML 문서 파싱 / DOM 트리 생성
2-3. DOM 트리 노드 탐색 / 스타스키마 인스턴스 추출
2-4. 데이터 웨어하우스에 인스턴스 삽입

본 장에서는 DTD로부터 스타 스키마를 설계하는 과정에 대해 설명한다. DTD로부터 스타 스키마를 설계하는 과정은 크게 4단계로 이루어진다. 먼저, DTD로부터 DTD 그래프를 생성한 후 다시 DTD 그래프로부터 애트리뷰트 트리를 생성한다. Attach/Graft/Prune과 같은 설계자의 작업을 거친 후 생성된 애트리뷰트 트리로부터 스타 스키마를 정의한다. 본 논문에서 고려하는 XML DTD는 그림 1과 같이 링크로 연결된 3개의 DTD로 구성되며, IDREF(S) 특성을 통해 ID 특성을 참조하는 특징을 포함하고 있다. 이러한 DTD로부터 스타 스키마를 설계한다.

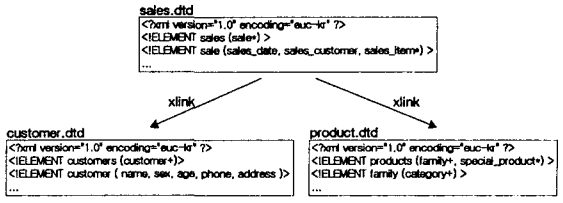

그림 1. XML DTD

3.1 DTD 그래프 생성

각 DTD의 엘리먼트와 애트리뷰트를 노드로 하는 DTD 그래프를 표 2와 같은 표기법을 통해 생성한다. 각 노드사이의 1:1 및 1:N 관계가 표시되고, 링크 및 IDREF 참조 관계가 그래프에 표시된다. 그림 2는 생성된 DTD 그래프 이다.

표 2. DTD 그래프 표기법

Notation	syntax	element	attribute
⊢→	1:N relationship	+, *	
→	1:1 relationship	default	#REQUIRED #FIXED
⊖→	optional	?, |	#IMPLIED
xlink →	link	XML 문서 사이의 xlink	
IDREF →	reference	element 사이의 IDREF(S)	

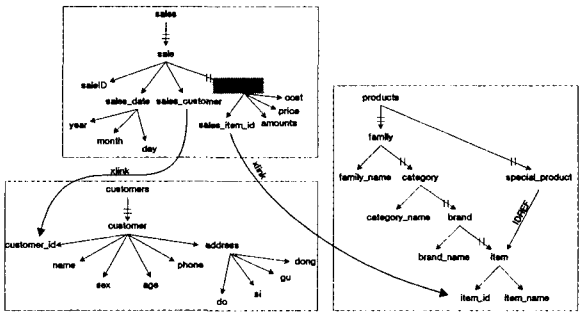

그림 2. DTD 그래프

3.2 애트리뷰트 트리 생성

DTD 그래프에서 애트리뷰트 트리를 생성하는 과정은 다음 표 3과 같다.

표 3. 애트리뷰트 트리 생성

1. 사실 지정 : 사실로 지정된 노드가 트리의 루트가 된다.
2. 루트로부터 단말 노드에 이르는 간선은 다음과 같은 관계를 통해 생성된다.
2-1. DTD 그래프에서 부모/자식 노드 사이의 N:1 relationship
2-2. DTD 그래프에서 부모/자식 노드 사이의 1:1 relationship
2-3. DTD 그래프에서 부모/자식 노드 사이의 optional relationship
2-4. DTD 그래프에서 노드 사이의 링크 (xlink →)
2-5. DTD 그래프에서 엘리먼트 사이의 IDREF (← IDREF)

3.3 Attach/Graft/Prune

생성된 애트리뷰트 트리에서 불필요한 노드를 제거하거나 필요한 노드를 추가(attach)하는 작업으로써 설계자의 작업에 해당된다. 일반적으로 DTD 그래프에서 데이터를 가지지 않는 노드가 그 대상이 되며, 다차원 분석에 직접적으로 관계되지 않는 노드를 설계자의 경험에 의해 제거(prune)하게 된다. 임의의 노드가 제거되면 그 노드의 자식 노드가

부모 노드와 연결(graft) 된다. 그림 3은 Attach/Graft/Prune을 거친 애트리뷰트 트리 이다.

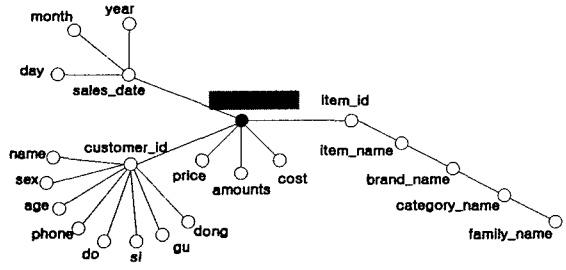

그림 3. 애트리뷰트 트리

3.4 스타 스키마 정의

스타 스키마 설계의 마지막 과정으로써 생성된 애트리뷰트 트리로부터 스타 스키마의 구성 요소를 정의한다.

1. 사실 테이블 및 측정값 정의
   그림 3의 애트리뷰트 트리에서 측정값은 amounts, price, cost가 되며, sales_date, customer_id, item_id를 외래키로 가지고 있다.
2. 차원 테이블 정의
   애트리뷰트 트리에서 차원은 루트 노드의 서브트리로 정의한다. 그림 3에서는 sales_date, customer_id, item_id를 루트로 하는 서브트리가 차원이 되고, 차원 테이블은 서브트리의 루트를 기본키로 하고 나머지 노드들을 애트리뷰트로 포함한다.
3. 차원의 계층구조 정의
   차원의 계층구조는 두 가지 형태로 정의된다. 먼저, 트리에서 노드 사이의 부모/자식 관계를 차원의 계층구조로 정의하고, 형제 노드들 사이의 관계에서는 구성원 수에 대한 정보를 통해 계층구조를 정의한다. 구성원수가 적은 노드가 구성원 수가 많은 노드의 상위레벨로 정의된다. 그림 3의 애트리뷰트 트리에서 정의되는 차원의 계층구조가 그림 4에서 보여진다.

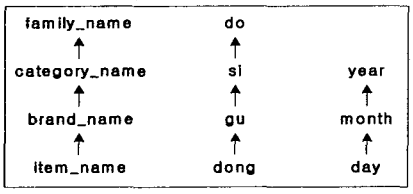

그림 4. 차원의 계층구조

4. 데이터 웨어하우스에 XML 문서 저장

그림 5는 지금까지의 설계 과정을 통해 설계된 스타 스키마를 보여주고 있으며, 이 스타 스키마에 대한 메타 정보를 이용하여 XML 문서로부터 인스턴스를 추출하여 관계형 데이터베이스에 저장한다.

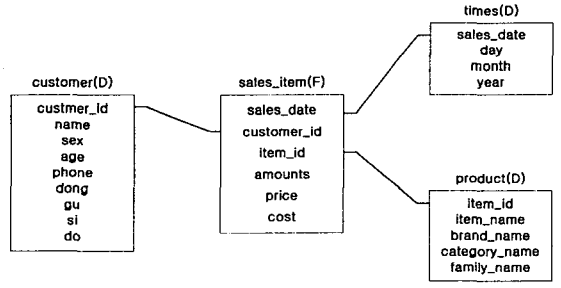

그림 5. 스타 스키마

XML 문서를 데이터 웨어하우스에 저장하기 위한 과정은 다음의 4단계로 이루어진다.

1. 설계된 스타 스키마의 메타 정보를 이용하여 데이터 웨어하우스에 스타 스키마를 정의한다.
2. XML 문서를 파싱하여 DOM 트리를 얻는다.
3. DOM 트리의 노드를 탐색하여 스타 스키마의 인스턴스를 추출한다.
4. 추출한 인스턴스를 데이터 웨어하우스에 삽입한다.

기업 및 사용자는 이렇게 생성된 스타 스키마를 OLAP 도구를 사용하여 다차원 분석에 이용할 수 있다.

5. 결 론

데이터 웨어하우스는 기업의 의사 결정 지원을 위해 OLAP에서 사용할 정보를 유지하고 있는 데이터의 집합체이다. 이 데이터 웨어하우스를 설계하기 위해서는 소스 데이터를 다차원 분석에 적합한 데이터 모델, 즉 스타 스키마로 변환해야 한다. 지금까지는 일반적으로 ER 도형으로 설계되어 있는 관계형 데이터베이스의 소스 데이터로부터 스타 스키마를 설계하는 연구가 진행되어 왔으나, 최근 인터넷의 발전으로 B2B, e-Commerce 등이 활성화되고, 차세대 웹 문서의 표준인 XML을 통한 정보 교환이 급증함에 따라 XML 문서에 대한 분석 요구가 생겨났으며, 따라서 데이터 웨어하우스 설계자는 XML 소스 데이터로부터 스타 스키마를 설계하는 작업을 수행해야 한다.

본 논문에서는 데이터 웨어하우스의 소스 데이터로써 XML 문서에 대한 스타 스키마를 생성하였다. XML DTD로부터 애트리뷰트 트리를 생성하여 스타 스키마를 설계하고, 설계된 스타 스키마의 메타 정보를 이용하여 XML 문서를 관계형 데이터베이스에 저장하기 위한 XML2Star 알고리즘을 개발하였다. 이를 통해 데이터 웨어하우스 설계자는 DTD를 통해서 XML 문서에 대한 스타 스키마를 설계하고, 이 DTD를 따르는 XML 문서로부터 스타 스키마의 인스턴스를 추출하여 데이터 웨어하우스에 저장할 수 있으며, 이것을 통해 기업 및 사용자는 OLAP 도구를 사용하여 XML 문서에 대한 다차원적인 분석을 수행할 수 있다.

본 논문에서 개발한 XML2Star 알고리즘에 대한 구현이 현재 진행중에 있다. 또한, DTD로부터 독립적인 잘 구성된(Well-Formed) XML 문서로부터 스타 스키마를 설계하는 연구가 향후 연구 과제로 남아 있으며, 관계형 데이터베이스의 소스 데이터와 XML 소스 데이터를 통합하여 스타 스키마를 생성하는 것이 향후 연구 과제로 남아 있다.

6. 참고 문헌

[1] S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology," Proc. of ACM SIGMOD Conf., pp. 65-74, 1997.

[2] 조재희, 박성진, "OLAP 테크놀로지," 시그마 컨설팅 그룹, 1999.

[3] L. Cabibbo and R. Torlone, "A Logical Approach to Multidimensional Database," Proc. of EDBT, pp. 183-197, 1998.

[4] Extensible Markup Language(XML) 1.0, http://www.w3.org/TR/PR-xml-971208.

[5] M. Golfarelli and S. Rizzi, "A Methodological Framework for Data Warehouse Design," Proc. of DOLAP, pp. 3-9, 1998.

[6] M. Golfarelli, D. Maio, and S. Rizzi, "Conceptual Design of Data Warehouses from E/R Schemes," Proc. of HICSS, pp. 334-343, 1998.

[7] M. R. Jensen, T. H. Moller and T. B. Pedersen, "Converting XML Data To UML Diagrams For Conceptual Data Integration," Proc. of DIWeb, pp. 17-31, 2001.

[8] M. R. Jensen, T. H. Moller and T. B. Pedersen, "Specifying OLAP Cubes On XML Data," Proc. of SSDBM, pp. 101-112, 2001.

[9] M. Golfarelli, S. Rizzi and B. Vrdoljak, "Data warehouse design from XML sources," Proc. of DOLAP, pp. 40-47, 2001.