지역 스키마간 충돌 문제를 고려한 XML 문서의 전역 스키마 생성 모델

김정희[O] 곽호영

제주대학교 통신컴퓨터공학부

{carina[O], kwak}@cheju.cheju.ac.kr

XML Global Schema Generation Model of XML Documents Considering Conflicts on Local Schema Integration

Jeong-Hee Kim[O] Ho-Young Kwak

Dept. of Communication & Computer Engineering, cheju National University

요 약

본 논문에서는 분산된 XML 문서를 수집 처리하여 상호 제공하는 과정에서 활용될 통합된 XML 문서의 전역 스키마를 생성하는 시스템을 제안한다. 제안된 시스템에서는 분산 환경에 존재하는 개별적인 XML 문서들에 대한 지역 스키마 정보를 관계형 데이터베이스로 구축하고, 통합된 XML 문서의 데이터를 기반으로 각각의 지역 스키마 데이터베이스를 검색한 후 데이터에 적합한 스키마 정의를 추출하게 된다. 또한 추출과정에서 중복 정의에 의한 충돌 범주를 분석하고, 이를 해결하는 방법도 제시하였다. 시스템 모델링 결과 XML 문서의 구조와 검증을 명확하게 보장하는 통합된 XML 문서의 전역 스키마 생성과 지역 스키마간 발생되는 충돌문제 해결이 가능함을 알 수 있었다.

1. 서 론

XML DTD(Document Type Definition)를 스키마 정보로 활용하는 데이터 통합(Data Integration) 방법 [1,2]은 인터넷상에서 다양한 이질 정보의 데이터들을 구조화된 문서로 표현하기 위한 언어인 XML(eXtensible Markup Language)이라는 형태로 통합하였다. 그럼으로써, 사용자들은 데이터가 저장된 시스템의 환경이나 저장형태와는 무관하게 동일한 질의를 사용하여 데이터들을 통합 검색할 수 있는 장점이 되었다. 하지만 DTD를 사용한 통합 방법은 DTD가 표현할 수 있는 데이터 타입의 제한과 XML 문서를 위한 파서(parser) 이외에 DTD 파싱(parsing)을 위한 별도의 파서가 필요하다[2].

최근 W3C에서는 기존 DTD의 단점을 보완한 XML 스키마 [3,4,5]라는 새로운 스키마 정의 언어를 제시하였으며, 이것은 XML 데이터에 대한 좀 더 정확한 표현을 위해 기존 DTD에서 표현할 수 있는 모든 데이터 타입들과 정수, 부동소수점, 사용자 정의형 데이터 등 다양한 데이터 타입들을 제공한다. 또한 XML 스키마는 그 자체가 XML의 한 응용이기 때문에 XML 스키마를 다루는 응용 프로그램에서는 별도의 파서가 필요치 않으며 기존의 XML 파서를 이용할 수 있는 장점을 가지게 되었다[6]. 하지만 이러한 XML 스키마의 장점에도 불구하고 스키마 통합 시 DTD에 비해 더욱 복잡하고 방대해진 데이터 타입과 제약조건(constraint)등으로 인해 XML 데이터에 대한 스키마 자동 생성과 기존 DTD를 스키마로 변환하는 등의 연구를 거쳐 데이터와 스키마에 대한 통합 관련 연구의 필요성을 대두시키고 있다[7].

따라서, 본 논문에서는 XML 관련 사업자들이 분산된 XML 문서를 수집 처리하여 상호 제공하는 과정에서 활용될 통합된 XML 문서의 전역 스키마 생성 과정을 모델링 한다. 이를 위해 분산된 XML 문서에 대한 지역 스키마 정보를 관계형 데이터베이스로 구축한 후 통합된 XML 문서를 스캐닝 하면서 파악된 정보의 스키마간 충돌 문제 해결 및 XML 문서 구조에 적절한 정의를 지역 스키마에서 추출하여 전역 스키마를 생성하도록 한다.

본 논문의 구성은 다음과 같다. 2장은 관련연구에 대하여 살펴보며, 3장은 XML 데이터의 지역 스키마 정보와 전역 스키마를 생성하는 방법, 그리고 지역 스키마간 충돌 문제의 분류와 해결방법을 기술하고, 4장에서는 전역 스키마 생성 시스템의 구조와 알고리즘을 제안한다. 그리고 마지막 5장에서는 결론 및 향후 연구과제를 기술한다.

2. 통합된 XML 문서의 전역 스키마 생성 모델

분산된 XML 문서를 수집하여 하나의 새로운 정보로 통합하는 과정에서 각각의 XML 문서를 정의하는 스키마를 지역스키마(local schema)라 하고, 통합된 XML 문서를 지원하는 스키마를 전역스키마(global schema)로 정의한다.

2.1 지역 스키마 정보 구축

이는 분산된 지역 스키마 정의를 따르는 XML 문서들이 통합되어서 하나의 새로운 XML 문서가 생성되었을 때 이 문서의 타당성과 구조를 검증하는 전역 스키마 생성 시 사용되며 구성되는 정보들은 네임 스페이스, 버전, 요소, 속성, 단순타입, 복잡 타입, 선택, 순차, 모든 그룹을 나타내는 요소 등의 다양한 옵션 정보들이며 관계형 테이블로 구축된다.

2.1.1 네임 스페이스 정보

XML 문서의 네임 스페이스와 버전 정보를 갖는다. 네임 스페이스는 XML 어휘집에 대한 레퍼런스를 가지는데, 기본 XML 스키마 요소들을 포함하는 속성과 표준 XML 스키마 데이터타입을 정의하는 속성을 갖는다. 그리고 네임 스페이스 접두사가 사용될 수 있다. 또한 버전 정보를 갖는다. 정보를 저장하기 위한 구조는 표 1의 ①과 같이 스키마 서두의 <schema> 요소의 속성 이름과 속성 값을 저장하도록 한다.

2.1.2 내용 모델 정보

내용 모델은 일반적으로 XML 문서에서 스키마의 서두(preamble) 구성이 끝난 후 기술되는데, 정해진 특정한 순서대로 나타나지 않는다는 특성을 갖는다. 따라서 네임 스페이스 테이블 정보가 구성된 후, 그 다음 나타나는 스키마 정의부터

스키마 정의 끝 부분까지 스캔하여 속성, 요소, 복잡 타입, any, group, 속성 그룹, 노테이션, 주석, include 정보들을 구축한다. 내용 모델 정보는 요소 테이블, 복잡 타입 테이블, 복잡 타입 내용 테이블로 구성한다.

▶ 요소 테이블

요소 테이블은 루트 요소와 자식 요소들의 정보로 구성된다. 요소(루트요소, 자식요소)들의 일반적인 형식은 <element> 키워드와 함께 이름(name)과 내용 모델을 기술하는 단순 또는 복잡 타입을 나타내는 형식(type), 그리고 기본 요소 내용과 고정 요소 내용, 그리고 널(Null)값 지정 여부, 카디널리티, 그리고 자신의 요소 ID 또는 자식 요소 ID 정보가 포함되도록 표 1의 ②와 같이 구성한다. 그리고 요소 테이블에서 참조되는 카디널리티(표 1의 ③), 기본 속성(표 1의 ④), 선택과 순차 등의 추가 속성들에 대한 정보(표 1의 ⑤)들도 테이블로 구축한다.

▷ 카디널리티 테이블은 요소 선언에서만 사용할 수 있는 옵션인 카디널리티 연산자 속성과 요소 테이블의 카디널리티ID를 테이블의 구성 요소로 정한다.

▷ 속성 테이블은 <attribute>를 사용한 속성 선언이나, 요소 내에서 선언된 속성들에 대한 정보를 구성한다. 특히 <attribute>를 사용한 선언은 다른 내용 모델에서도 사용할 수 있기 때문에 ref 정보도 구성하도록 한다.

▷ 선택, 순차, 모든 그룹 테이블은 요소들이 나타나는 순서나 임의 택일, 그리고 리스트화하는 정보들로 구성된다. 값으로는 sequence, choice, all을 갖는다.

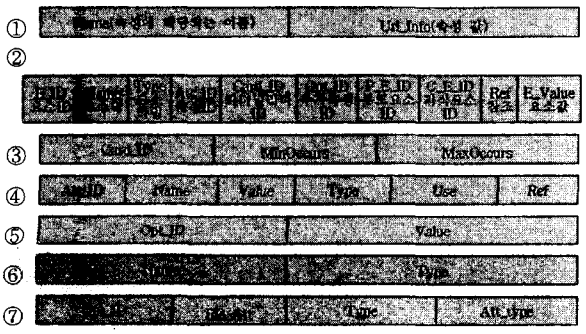
▶ 복잡 타입 테이블

복잡 타입은 <complexType> 요소와 그 속성 및 적합한 구속 패싯으로 정의되는데, 요소 선언, 속성 선언 및 요소 참조를 포함한다. 일반적으로 복잡 타입을 사용하는 요소들이 선언될 때 형식의 속성 값으로 정의가 되므로, 여기서 Name은 해당 요소의 이름이며, Type은 이 Name의 요소가 포함될 타입 정의를 결합시키는데 사용되도록 표 1의 ⑥과 같은 구조로 구성한다. 그리고 해당되는 Type에 대한 세부 내용은 복잡 타입 내용 테이블에 구축한다.

▶ 복잡 타입 내용 테이블

복잡 타입 내용 테이블은 위에서 기술된 복잡 타입 테이블의 각 요소의 Type들에 대한 세부 선언들로 구성된다. 복잡 타입의 이름과 추가 속성, 그리고 포함되는 요소와 그에 대한 속성들로 정보가 구성되는데 요소 테이블 정보와 혼합하지 않고 표 1의 ⑦과 같은 구조로 독립되도록 구성한다. 이는 복잡 타입이 하나의 객체로 취급되기 때문이다. 표 1은 지역 스키마 정보를 구축하기 위한 데이터베이스 스키마이다.

표 1. 지역 스키마 데이터베이스 모델



2.2 전역 스키마 정보

전역 스키마는 어떤 하나의 XML 문서가 분산 되어있던 개별적인 XML문서에서 추출되었을 때 참조하게 되는 스키마이다. 인터넷의 특징인 분산 환경에서 업무를 처리하다 보면, 분산된 여러 곳의 정보를 참고하여 하나의 통합된 정보를 구축해야 하는 필요성이 존재한다. 통합된 정보는 XML 문서이며, 따라서 이 XML 문서의 구조를 정의하고 검증하는 스키마의 존재가 필수조건이 된다. 개별적인 XML 스키마는 2.1절에서 기술한 방식으로 정보들이 구축되었으며, 전역 스키마는 통합된 XML 문서를 스캔하여, 사용되어지고 있는 요소, 속성들을 해당되는 지역 스키마에서 참조하고 추출하여 전역 스키마를 구성한다.

2.3 충돌 문제와 해결

전역 스키마 생성 시 통합된 XML 문서의 내용이 지역스키마에 중복 존재하는 경우, 스키마 충돌 문제가 발생된다. 즉, 어느 쪽 지역 스키마를 참조해야 하는지 결정을 해야하는데, 이는 대부분 통합된 XML 문서에서 값을 나타내는 요소나 속성들에 제한된다. 따라서, 충돌이 발생되면 현재 값의 속성을 참고하여 값을 정의하는 지역 스키마에서 해당되는 정의를 전역 스키마로 삽입한다. XML 지역스키마간 스키마 충돌과 의미 충돌 문제의 경우 [6]에서 분류한 기준을 따른다.

2.3.1 naming conflict

▷ synonym

<subject title="my schedule"> 와 같이 통합된 XML 문서의 "subject" 요소가 지역스키마에 중복 정의되었을 때 "subject" 요소의 속성 값을 참고하여 속성 값에 해당되는 데이터 타입이 정의된 지역 스키마의 정의를 추출하여 전역 스키마에 삽입한다.

▷ homonym

지역스키마에 각각 "ProductName"과 "PName"이 서로 이름은 다르지만 의미가 동일한 경우 어느 한쪽의 지역 스키마를 사용할 수도 있고, 다른 쪽 지역 스키마를 사용할 수도 있다. 하지만, 어느 한쪽의 지역스키마를 사용하게되면 동일 데이터를 서로 다르게 처리하게 되는 문제가 발생되므로 전역 스키마를 구성할 때 새로운 요소를 정의하고 이 요소에 지역스키마 정보를 포함하도록 한다.

2.3.2 structural conflict

구조적 충돌은 이름은 동일하지만 타입이 상이한 *<part_num id="Global001">* 와 *<part_num id="1">* 와 같은 경우이며 동일한 타입이면서 요소 또는 속성은 *<part_no id="Global001">* 와 *<part_no id="Global100">* 처럼 이름과 속성이 동일하지만 part_no는 요소로 정의되어 있고, 다른 쪽 part_no는 속성으로 정의되는 경우이다. 모두 통합된 XML 문서의 속성 값을 참고하여 속성 또는 요소 값에 해당되는 데이터 타입이 정의된 지역 스키마의 정의를 추출하여 전역 스키마에 삽입한다.

2.3.3 type conflict

타입 충돌은 기본형과 기본형, 기본형과 유도형, 유도형과 유도형 간의 충돌이 있다. 하지만 타입 충돌의 해결은 3개의 충돌 범주보다는 값을 기준으로 서로 값이 상이하다면 이는 통합된 XML 문서의 값을 참고하여 해당되는 값을 정의한 지역 스키마를 사용해야 하며, 서로 값이 동일 타입이면 widening 변환을 하도록 한다.

2.3.4 단위 충돌, 포함성 단위 충돌, 표현 충돌

단위 충돌은 <price>10</price>에서 "10"이라는 값이 서로 다른 단위일 때, 그리고 포함성 단위 충돌은<sell> ?? </sell>이 단위 충돌에 포함되는 요소 안의 서로 다른 범위의 값을 표현하는 경우에 해당된다. 그리고 표현 충돌은 같은 요소의 값이 서로 같은 의미면서 다른 표현인 경우이다. 표현 충돌은 위에서 언급된 homonym 처리 방법을 이용한다. 단 동일 요소 이름이므로 요소 이름에 대한 재정의 없이 단지 관련되는 참고 정보인 <import> 요소만 추가하도록 한다. 그리고 단위 충돌과 포함성 단위 충돌은 서로 포함성 관계가 있고, A와 B 시스템 중 어느 한쪽의 정의를 따르면 데이터의 불일치가 발생된다. 상호간 변환 Tool이나 주석을 이용하는 해결방법이 가능하다.

3. 스키마 생성 시스템 구조 및 설계

3.1 지역 스키마 데이터베이스 생성기

분산된 XML 문서의 스키마를 스캐닝 하여 2장에서 기술한 데이터베이스를 생성한다. 생성된 데이터베이스는 전역 스키마 생성 시 전역 스키마 생성기에 의해 사용된다. 알고리즘은 그림 1과 같다.

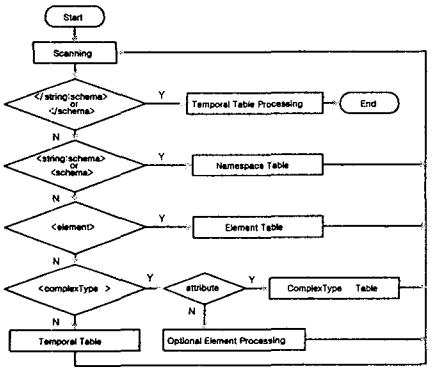


그림 1. 지역 스키마 데이터베이스 생성 알고리즘

3.2 전역 스키마 생성기

XML 문서의 구조를 정의하고, 또한 검증하기 위해 사용되는 스키마를 생성하기 위해 통합된 XML 문서를 스캐닝 하면서 파악되는 정보를 기준으로 해당 지역 스키마 데이터베이스를 이용하여 통합된 XML 문서의 전역 스키마를 생성한다. 알고리즘은 그림 2와 같다.

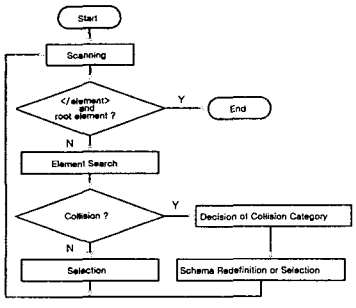


그림 2. 전역 스키마 생성 알고리즘

3.3 시스템 구조

통합된 XML 문서를 위해 지역 스키마 데이터베이스 생성기와 전역 스키마 생성기를 모듈로 전역 스키마 생성 시스템의 전체 구조는 그림 3과 같다.

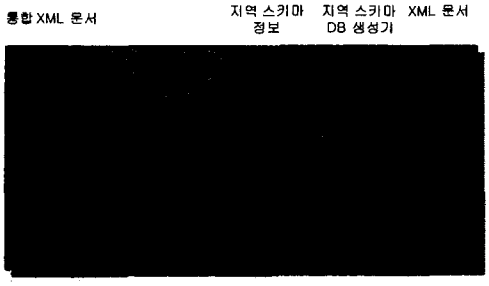


그림 3. 전역 스키마 생성 시스템

4. 결 론

본 논문에서는 분산 환경에 존재하는 개별적인 XML 문서들을 기반으로 통합된 하나의 XML 문서를 생성하였을 때 이 XML 문서의 구조를 정의할 수 있으며, 또한 검증하는데 필수 조건인 XML 전역 스키마를 생성하는 시스템을 제안하였으며, 또한 지역 스키마간 중복 정의에 의한 충돌 문제에 대해서도 충돌 범주에 따라 그 해결 방법을 제시하였다. 하지만 제안된 시스템은 통합된 XML 문서에 대한 전역 스키마 생성 및 지역 스키마간 충돌 문제의 해결 가능함을 보여주지만 지역 스키마에 대한 정보가 관계형 데이터베이스로 구축됨에 따라 전역 스키마 생성 시 데이터베이스 handling 이라는 작업이 필요하게 되었다. 따라서 향후 연구는 본 시스템의 구현과 또한 XML 스키마가 XML 문서라는 장점을 활용하여 파서(parser)만을 이용한 전역 스키마를 생성하고 두 시스템을 비교·분석 하고자 한다.

참고문헌

[1] 이강찬, 이경하, 이규철, "XML 기반의 인터넷 정보 자원 통합", 데이터베이스 연구, 한국데이타베이스연구회, 제16권 2호, pp. 5-21, 2000. 12.

[2] C. Baru, A Gupta, B. Ludascher, R. Marciano, Y. Papakonstantinou, P. Velikhov, V. Chu, "XML-Based Information Mediation with MIX", exhibition program, ACM Conf. on SIGMOD'99, Philadelphia, USA

[3] W3, Recommendation, "XML Schema Part 0: Primer", http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/, 2001

[4] W3, Recommendation, "XML Schema Part 1: Structure", http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/, 2001

[5] W3, Recommendation, "XML Schema Part 2: Datatypes", http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/, 2001

[6] 이승원, 권석훈, 김미혜, 이경하, 이규철, "XML Schema를 이용한 스키마 통합시 충돌 문제의 분류", 정보과학회 학술대회 Vol. 28, No. 2, pp. 31-33, 2001. 10

[7] Behrens, R.: A Grammar Based Model for XML Schema Integration; in: Lings,B. et al (Eds.): Advances in Databases, 17th British National Conference on Databases, London, LNCS, Vol. 1832, S. pp. 172-190, 2000. 7.