XML Schema를 이용한 스키마 통합에서 충돌 문제의 포괄적 분류와 해결

박현미[0] 박 석
서강대학교 컴퓨터학과
(mi2you[0], spark)@dblab.sogang.ac.kr

Comprehensive Classfication and Resolution of schema conflicts on XML Schema Integration

Hyun mi Park[0], Seog Park
Dept. of Computer Science, Sogang University

요 약

웹을 통해 얻을 수 있는 데이터의 양이 방대해 지면서, 이질적이고 분산된 웹상의 데이터들을 통합하여 이용하려는 요구가 커지고 있다. 데이터의 통합은, 각 데이터의 스키마를 통합하여 단일화된 스키마를 만들고 통합된 스키마에 질의를 하여 원하는 결과를 얻는 것으로 이룰 수 있다. 이러한 스키마 통합의 필요성은 인터넷 환경이 보편화되고 정보의 양이 방대해 지면서 웹 데이터를 대상으로 하여 더욱 커지게 되었다. 본 논문에서는 XML의 새로운 스키마 언어인 XML Schema 를 이용한 스키마 통합시에 발생하는 스키마 충돌의 포괄적인 분류와, 이때 발생하는 충돌을 해결하고 통합 스키마를 작성하기 위한 기법을 제안한다.

1. 서 론

전통적인 데이터베이스 시스템에서 스키마 통합의 목적은 여러 개의 데이터베이스를 마치 하나처럼 사용자에게 보여주기 위한 것이다. 통합된 뷰인 전역 스키마(global schema)를 제공하여 여러 지역 소스에 직접 접근하지 않도록 사용자 편이성을 증가시키고 데이터의 투명성을 증가시킨다. 본 논문은 웹데이터의 표준인 XML 통합을 위해 스키마의 통합을 고려한다. 통합시 각 정보 자원의 특성에서 기인하는 다양한 이질성에 의해 스키마 간의 충돌이 발생할 수 있으며, 이러한 충돌 문제는 정보자원 통합에 있어서 해결해야 할 가장 중요한 문제 중의 하나이다. 예를들어, [학부생]을 표현한 독립적으로 디자인된 (a), (b) 두개의 지역스키마에서 (a)의 학부생은 "학부생"이라는 이름으로, (b)의 학부생은 "학생"이라는 이름으로 표현될 수 있다. (a)의 학부생은 이름,학번,학과를 속성으로 가지는 데 반해, (b)의 학부생은 이름, 학과, 주소를 속성으로 가질 수도 있다. 하나의 스키마로의 통합은 그리 간단한 문제가 아니다. 기존의 XML의 스키마 언어는 주로 DTD가 사용되었으나 여러 한계점으로 인해 W3C는 새로운 스키마 언어인 XML Schema[4]를 표준화하였다. 또한, 기존의 연구에서 DTD는 주로 문서의 유효성 검증을 위해 사용되었으나, XML Schema는 기존 DBMS의 스키마처럼 모델링의 가능성을 포함하면서, 데이터의 '개념'에 대한 정보도 내포하고 있다. 본 논문에서는, 스키마의 통합시 해결해야 하는 스키마 충돌문제에 관한 포괄적인 분류를 통하여, 통합할 스키마간의 충돌을 찾고, 발생한 충돌 문제를 각각의 해결방법을 통해 해결하여 통합 전역 스키마(global schema)를 얻기 위한 기법을 제안한다.

2. 관련 연구

정보통합을 위한 스키마 통합의 연구는 오래 전부터 데이터베이스 분야에서 연구되어 왔다. 이질적인 데이터베이스를 대상으로 스키마를 통합하는 방법은 멀티데이터베이스에서 집중적으로 연구되었으며, 이 때 발생하는 스키마 충돌의 분류와 해결방법도 제안되었다[2,3]. 그러나, 멀티데이터베이스에서의 이러한 분류는, 통합의 대상이 DBMS로 한정되어 있으며 통합 데이터 모델 또한 RDB와 OODB가 지원하는 데이터모델을 이용한다.

XML을 위해 스키마 통합을 시도한 예는 MIX[5]와 XMF[6]에서 찾아볼 수 있다. 두 시스템 모두 DTD를 XML의 스키마로 사용한다. MIX는 XMAS 모듈을 이용하여 미디에이터 뷰를 정의하고, DTD Inference 모듈이 주어진 미디에이터의 뷰 정의와 소스 DTD를 가지고 view DTD를 얻어낸다. Tightest DTD를 생성해내는 알고리즘을 통해, 모든 지역 소스에 포함되는 가장 제약된 형태의 view DTD를 얻는다. 이는 단순히 문법상의 통합이다. XMF에서는 미디에이터 어드민이 지역DTD를 참고하여 적합한 전역 DTD를 작성하고, 충돌의 해결을 위해서 전역 스키마와 지역 스키마간의 연관관계를 기술한 중재규칙을 이용한다.

여러 개의 스키마가 통합될 때 각 요소의 의미에 대한 정확한 정의는 성공적인 통합을 위해 중요하다. "의미"에 대한 정확한 기준과 이해를 제공하기 위해 온톨로지를 이용할 수 있다.

3. 스키마 통합을 위한 시스템 구조와 절차

웹 데이터를 대상으로 통합시스템을 구축할 때 가장 적합한 시스템 환경으로 고려되는 것은 래퍼-미디에이터 시스템이다. 그림 1 은 미디에이터 환경에서 XML Schema를 이용하여 통합 스키마를 얻기 위한 시스템 구조도이다. 래퍼는 통합하고자 하는 정보자원의 데이터를 미디에이터가 중재할 수 있도록 공통 데이터 모델로 변환하는 작업을 수행하는데, 이때 공통 데이터 모델로는 XML을 사용하고 스키마로 XML Schema를 사용한다. 통합(Integration) 모듈을 통해 지역 XML Schema들을 비교하고, 충돌을 찾아내어 해결한다. 충돌 해결을 지원하기 위해

메타데이터(metadata)을 이용하게 되는데, 의미적 기준과 함께 공통 어휘집을 제공하는 도메인 온톨로지(domain ontology), 통합 스키마와 지역스키마 간의 매핑정보, 그리고 통합스키마와 지역스키마간의 매핑을 지원하는 함수들로 구성된다.

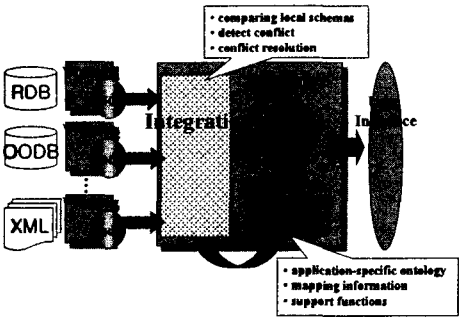

그림 1. XML Schema를 이용한 통합 시스템 구조도

통합 스키마를 얻기 위해, 1)전처리 통합, 2)비교, 3)통합 의 3단계의 통합 절차를 거친다.

[단계1] 전처리 통합 (Pre-integration)

통합 할 지를 결정하기 전에 스키마를 분석하고, 통합을 위한 정책들을 결정하는 단계이다. XML Schema의 통합을 위해 두 가지 정책을 갖는다.

[단계2] 비교 (Comparing)

스키마간의 이질성으로 인해 발생하는 충돌을 찾아낸다. XML Schema를 이용한 통합시 발생할 수 있는 스키마 충돌은 충돌의 종류와 대상에 따라 달라진다. 충돌의 자세한 분류는 4장에서 설명한다.

[단계3] 통합 (Integration)

스키마간의 비교를 통해 발견된 충돌을 해결하고 통합 스키마를 작성한다. 통합 스키마가 가져야 할 3가지 조건은 완전성, 최소성, 이해성이다.

4. XML Schema를 이용한 스키마 통합시 충돌분류

XML Schema를 대상으로 한 스키마 충돌의 종류를, 이름충돌, 타입충돌, 제약충돌로 나누었고, 각각의 의미는 다음과 같다.

- 이름충돌(Naming conflict) : 같은 개념을 다른 용어를 사용하여 표현한 경우(이음동의어), 혹은 다른 개념을 같은 용어를 사용하여 표현한 경우(동음이의어)에 발생한다.
- 타입충돌(Type conflict) : 의미적으로 같은 엘리먼트나 어트리뷰트가 서로 다른 구조나 데이터타입을 가지는 경우에 발생한다.
- 제약충돌(Constraints conflict) : 의미적으로 같은 엘리먼트나 어트리뷰트가 서로 다른 제약사항을 가지는 경우에 발생한다.

XML Schema에서 스키마 충돌의 대상은 크게 엘리먼트(element)와 어트리뷰트(attribute)로 나뉜다. 엘리먼트는 다시, 복합 엘리먼트와 단순 엘리먼트로 나누었다.

- 복합 엘리먼트 (complex element) : 내용모델(content model) 내부에 자식 엘리먼트나 어트리뷰트를 가지는 엘리먼트, 복합 타입(complexType)을 가지는 엘리먼트
- 단순 엘리먼트 (simple element) : 복합 엘리먼트가 아닌 엘리먼트, 단순 타입(simpleType)을 가지는 엘리먼트
- 어트리뷰트 (attribute)

그림 2 는 충돌의 종류와 대상에 따라 발생할 수 있는 모든 충돌의 분류를 하나의 표로 나타낸 것이다.

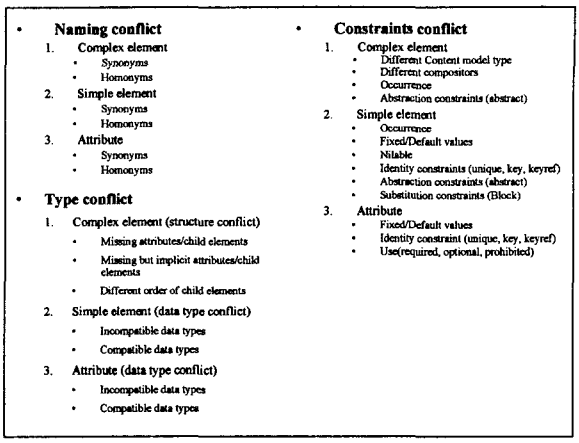

그림 2. XML Schema 통합시 스키마 충돌의 포괄적 분류

4.1. 이름충돌

이름충돌에는 두 가지 종류가 있는데 의미적으로 같은 엘리먼트/어트리뷰트에 다른 이름을 사용하는 경우(이음동의어)와 의미적으로 다른 엘리먼트/어트리뷰트에 같은 이름을 사용하는 경우(동음이의어)이다.

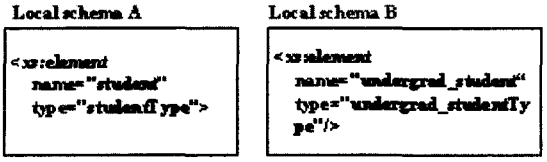

그림 3. 이름충돌 (이음동의어)

4.2 타입충돌

두개의 스키마에서 같은 개념을 표현하는 엘리먼트나 어트리뷰트가 서로 다른 타입을 가질 경우 발생하는 충돌이다. 타입 충돌은 다시 두 가지 종류로 나뉘는데, 복합 엘리먼트에 의한 구조타입충돌과, 단순 엘리먼트나 어트리뷰트에 의한 데이터타입충돌이다. 복합 엘리먼트의 타입은 어트리뷰트나 자식 엘리먼트를 포함하는 내용모델(content model)을 가진다. 내용모델의 차이로 인해 발생하는 이러한 충돌을 '구조타입충돌'이라 부른다. XML Schema는 44+ 개의 기본형(built-in type)을 가진다. 또한 데이터 타입별로 지원하는 다양한 파셋들을 통해 다양한 유도형(user-driven type)을 얻을 수 있으며, 이로 인해 발생하는 충돌을 '데이터타입충돌'이라 부른다.

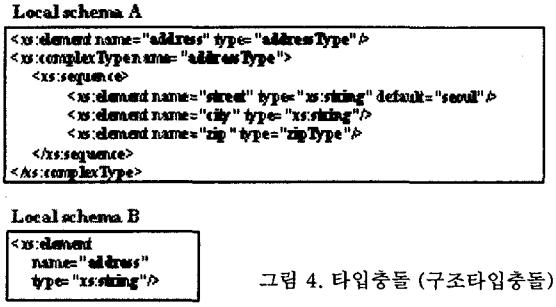

그림 4. 타입충돌 (구조타입충돌)

4.3. 제약충돌
두개의 스키마에서 같은 개념을 엘리먼트나 어트리뷰트 제약사항의 명세가 다를 경우에 발생한다. XML Schema는 구성자의 차이, 발생횟수의 차이, 기본값, 키필드설정 이외에도 매우 다양한 제약사항을 설정할 수 있고 이로 인해 충돌이 발생한다.

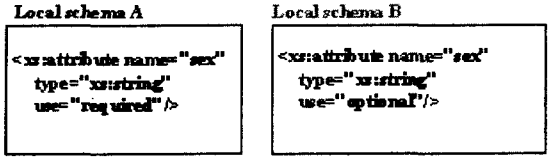

그림 5. 제약충돌 (use속성충돌)

5. 충돌의 해결과 통합 스키마 작성

스키마 통합의 3단계 중, 세번째 단계인 통합(integration)에서는, 발생한 스키마 충돌을 해결하고 통합 스키마를 작성한다. 통합 스키마의 조건인 "작성된 통합 스키마의 범위, 구조, 발생빈도등은 모든 지역스키마를 포함하거나 최소한 같아야 한다"를 만족하도록 통합스키마는 작성되어야 한다.

- 이름충돌해결 : 중앙에서 관리되는 하나의 카탈로그를 통해 통합 스키마와 지역 스키마의 이름사이의 매칭 정보를 저장하고, 재명명 방법을 통해 이름충돌은 해결할 수 있다. 온톨로지를 구성하는 용어들로 구성된 공통 어휘집(common domain specific vocabulary)을 카탈로그로 사용하여 재명명 함으로써 의미에 대한 공통의 이해를 돕는다.
- 타입충돌해결
  - 구조타입충돌해결 : 자식 엘리먼트와 어트리뷰트의 공통되는 부분을 먼저 선언하고, 한쪽 스키마에서만 찾을 수 있는 엘리먼트는 선택적(optional) 선언을 한다. 엘리먼트의 경우 minOccurs를 '0'으로 선언하고, 어트리뷰트의 경우 use 속성을 'optional'로 선언한다. 또한, Differrent order 충돌의 경우 <choice> 구성자를 이용하여, 조합 가능한 자식 엘리먼트의 순서 중 하나를 선택할 수 있는 통합 스키마를 구성함으로써 해결할 수 있다.
  - 데이터타입충돌해결 : 모순되는 데이터타입충돌해결은 <union>을 사용하여, 여러 개의 데이터타입을 동시에 선언할 수 있게 함으로써 해결할 수 있다. 호환되는 데이터타입충돌해결은 타입 자동변환을 통해 해결할 수도 있다. 그 예로, 문자형의 길이 차이, 패턴 차이, 숫자형의 범위차이 등을 들 수 있다.
- 제약충돌해결 : 이러한 충돌은 대부분 미디에이터 운영자가 통합 스키마의 조건에 만족하도록 제약조건을 지정함으로 해결한다.

충돌을 발견하고 해결하여 사용자에게 필요한 통합스키마를 만드는 과정은 완전히 자동화 되기는 어렵고, 미디에이터 운영자의 개입을 필요로 한다. 또한 제시한 해결방법은 4절에서 분류한 충돌문제중 일부에 대한 해결방안이다.

6. 평가

제안한 충돌분류는 발생하는 스키마 충돌을 종류별로 나누고 그 내부에 각 스키마 요소별 발생 가능한 충돌들을 분류하여, 어떠한 스키마 모델이라도 적용하기 쉬운 분류 형태를 가지고 있다. XML Schema를 사용하여 스키마 충돌을 분류한 예는 [7]에서도 찾아볼 수 있다. 그러나 [7]의 분류는 멀티데이터베이스에서의 스키마 충돌 분류를 XML Schema에 그대로 적용시켰다. 그러나, XML Schema와 DBMS의 스키마는 스키마 모델, 충돌대상, 스키마 정의 언어등의 차이점을 갖고 기존의 스키마에 비해 뛰어난 표현능력을 가지므로 기존의 분류를 그대로 적용하는 데는 무리가 따른다. 스키마 충돌 분류를 위한 평가는 완전성, 최소성, 정확성,이해성, 적합성의 기준을 가지고 평가하였다. 제안한 충돌 분류는 모든 가능한 충돌 요소들이 포함되었고 -완전성-, XML Schema 언어명세를 완벽하게 검토하여 정확성을 높였고-정확성-, 분류상에 중복을 최소화하였고-최소성-, 가장작은 분류단위를 사용하여 이해하기 쉽게 하였다-이해성-.

7. 결론 및 추후연구

본 논문에서 제안한 기법을 통해 기여한 사항은 다음과 같다.

1) XML의 새로운 스키마 표준인 XML Schema를 이용한 통합시에 발생하는 스키마 충돌문제의 포괄적인 분류를 함으로써, XML의 통합과 교환을 수행하는 어플리케이션들을 위한 기초 연구로서 의의를 가진다.
2) 충돌을 해결하고 통합 스키마를 만들기 위한 기법들을 제안하였다. 통합을 위한 각 단계를 통해, 충돌별 해결방안을 구체적을 제안하였다. 또한, 공통 개념 어휘집을 제공하는 온톨로지의 이용으로, 의미적으로 보다 정확하고 재사용 가능한 통합 스키마를 구축하였다.

충돌 해결을 위한 가장 적합한 전체적인 프레임웍에 관한 연구가 수행되어야 한다. 통합 스키마를 만드는 과정은 완전히 자동화되기는 어렵고, 미디에이터 운영자의 개입을 필요로 한다. 반자동화 시킬 수 있는 시스템의 개발로, 미디에이터 운영자의 개입을 최소화 하기 위한 방법에 대한 연구가 더 이루어져야 한다.

참고문헌

[1] C. Batini, M. Lenzerini, and S. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, Vol. 18, No. 4, Dec. 1986, pp. 323-364.

[2] Won Kim, Injun Choi, Sunit Gala, and Mark Scheevel, "On Resolving Schematic Heterogeneity in Multidatabase Systems," Modern Database Systems; The Object Model, Interoperability, and Beyond, Addison-Wesley Publishing Company, pp. 521-550, 1995.

[3] Won Kim and jungyun Seo, "Classifying Schematic and Data Heterogeneity in Multidatabase Systems", Computer, pp. 12-18, december 1991.

[4] W3C XML Schema. Available by WWW in: http://www.w3.org/XML/Schema.

[5] Yannis Papakonstantinou, Pavel Velikhov, "XML-Based Information Mediation with MIX", UCSD, 1999

[6] 이강찬, 이경하, 이규철, "XML 기반의 인터넷 정보 자원 통합", 한국정보과학회지, 16권 2호, pp. 5-21, 2000

[7] 이승원, 권석훈, 김미혜, 이경하, 이규철, "XML Schema를 이용한 스키마 통합시 충돌 문제의 분류", 2001 한국정보과학회 추계 학술발표회논문집(I), pp 31-33, 2001