

이질 정보 자원 통합을 위한 통합 스키마 저작 도구 설계¹⁾

김미혜*⁰ 이승원* 이경하* 이규철* 김병섭** 이미영**
 *충남대학교 컴퓨터공학과, **한국전자통신연구원 컴퓨터소프트웨어연구소 콘텐츠 기술연구부
 (mhkim⁰, swlee, bart, kcllee)@flower.ce.cnu.ac.kr, (powerkim, mylee)@etri.re.kr

An Implementation of Authoring Toolkit for Global Schema Definition on Heterogeneous Information Resources

Mi-Hye Kim⁰ Seung-Won Lee Kyong-Ha Lee Kyu-Chul Lee

Dept. of Computer Engineering, Chungnam National University

Byoung-Seob Kim Mi-Young Lee

Dept. of Contents Technology, Computer & Software Research Laboratory, Electronics and Telecommunications Research Institute

요 약

인터넷의 사용이 보편화되어 가고 있는 현대 정보 사회에서는 다양한 정보들이 인터넷 상에 분산되어 존재하고 있다. 이러한 정보 자원들은 서로 이질적인 형태로 분산되어 존재하기 때문에, 정보를 얻으려 하는 최종 사용자가 이러한 정보 자원들에 대해 단일화된 형태로 접근할 수 있는 방법이 요구된다. 따라서, 이러한 정보 통합에 대한 많은 연구가 진행되어 왔다. 본 논문에서는 XML Schema를 이용한 정보 통합에 있어서, 각 정보 자원들의 이질성에 기인하는 충돌 문제를 분류하고, XQuerySD를 이용한 뷰 정의 방식의 통합 과정에서 충돌 해결 방안을 제시하며, XQuerySD를 쉽게 생성할 수 있도록 도와주는 저작 도구를 설계한다.

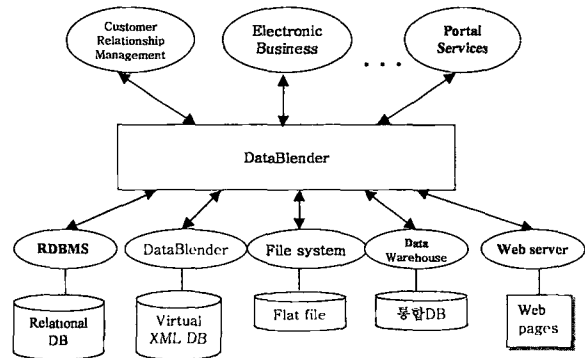
1. 서 론

인터넷의 사용이 보편화되어 가고 있는 현대 정보 사회에서는 다양한 정보들이 인터넷 상에 분산되어 존재하고 있다. 이러한 정보 자원들은 서로 이질적인 형태로 분산되어 존재하기 때문에, 정보를 얻으려 하는 최종 사용자가 이러한 정보 자원들에 대해 단일화된 형태로 접근할 수 있는 방법이 요구된다. 따라서, 이러한 정보 통합에 대한 많은 연구가 진행되어 왔다.

본 논문에서는 정보 통합 과정에서 발생하는 데이터 이질성을 해결하기 위한 공통 데이터 모델로 XML을 사용하고, XML 데이터에 대한 스키마 표현으로는 XML Schema를 사용함으로써 스키마 표현에 대한 이질성을 해결한다. 또한, 유사한 정보 자원들에 대해 서로 다른 스키마 정의에서 발생하는 이질성과 데이터의 의미적인 차이에서 발생하는 이질성을 분류하고 각각에 대해 해결 방안을 제시하며, 사용자가 충돌 문제를 쉽게 해결할 수 있도록 도와주는 저작 도구의 설계를 목적으로 한다.

본 논문에서 설계하는 저작 도구가 적용되는 목표 시스템인 DataBlender는 통합 자료 모델을 제공하고, 다양한 자료 저장소에 대한 래퍼를 제공하여 정보 통합을 실현하는 미디어기반 정보 통합 시스템으로, 인터넷 정보뿐만 아니라 인터넷 상에 분산되어 있는 다양한 정보들을 활용하여 전자상거래, 포털 서비스, 고객 관계 관리 시스템 등을 구축할 수 있도록 ETRI에서 설계 및 구현되고 있는 시스템이다.

DataBlender는 통합 데이터 모델로 XML을 지원하고, 통합 스키마 규격으로는 XML Schema를 지원한다. 그리고 각 지역 스키마로부터 통합 스키마를 정의하는 통합 스키마 및 중재 규칙 정의어로 XQuery 기반의 XQuerySD(XQuery for Schema Definition)를 제공하며, XQuery를 통해 통합 정보에 대한 검색을 지원한다.



[그림 2] DataBlender 시스템 구조

DataBlender에서 통합 스키마와 지역 스키마 정보를 표현하는 방법으로 사용하는 XML Schema[4]는 기존의 스키마 정의 언어인 DTD의 단점을 보완하기 위해 W3C에서 제안하는 스키마 정의 언어로, 44개 이상의 데이터 타입을 가지고, 기본 타입에 패싯(facet)을 적용함으로써 스키마 작성자가 데이터 타입을 스스로 정의할 수 있으며, 한 데이터 타입을 확장하거나 제약 조건을 줄 수 있다. 또한 XML Schema는 XML과 동일한 문법을 사용하기 때문에 스키마 기술을 위한 별도의 기술을 습득할 필요가 없다.

정보 통합 시스템에서는 통합 스키마와 지역 스키마 간의 대응 관계를 표현할 수 있는 스키마 정의 언어가 필요하다. 본 논문에서 제안하는 시스템에서는 이와 같은 요구에 의해 스키마 정의 언어인 XQuerySD[1]를 정의하였다.

XQuerySD는 XQuery에 기반한 통합 스키마 정의 언어로, 정의된 XQuerySD는 통합 시스템 내부에서 XML Schema 규격으로 변환되어 저장된다. XQuerySD는 XQuery의 거의 모든 기능과 특성을 가지고 있지만 XQuery가 실제 XML 데이터에 대한 질의문이라면 XQuerySD는 XML 데이터의 스키마 정보 등을 기술한 XML Schema에 대한 질의문이다. 이러한 XQuerySD는 그 결과값이 XML Schema 형태로 나타나기 때

1) 본 논문은 한국전자통신연구원 의 위탁연구과제인 이질 자료 모델의 충돌 해결 방안 연구(과제번호 : 104573)의 일부로 수행된 결과임

문에 실제 XML 데이터를 결과값으로 하는 XQuery에서 사용되는 모든 연산을 제공할 필요는 없다.

2. 충돌의 분류와 해결 방안

2.1 충돌의 분류

XML Schema를 이용한 스키마 통합에 있어서 발생할 수 있는 충돌 문제는 [표 1]과 같이 데이터들의 스키마 이질성에 의해 발생할 수 있는 스키마 충돌(Schematic Conflict)과 데이터 간의 의미적인 이질성에 의해 발생할 수 있는 의미적 충돌(Semantic Conflict)로 분류할 수 있다[1, 2].

[표 1] 충돌 대상에 따른 충돌 분류

| | |
|---|---|
| <ul style="list-style-type: none"> ■ 엘리먼트 대 엘리먼트 ■ 1 대 1 엘리먼트 <ul style="list-style-type: none"> ▶ 엘리먼트 이름 -동음이의 -이음봉의 ▶ 엘리먼트 제약조건 <ul style="list-style-type: none"> -반정빈도수(occurrence) -컨텐츠 모델(데이터 타입) -기본값 -널값 ▶ 엘리먼트 구조 <ul style="list-style-type: none"> -에트리뷰트 누락 -모호한 에트리뷰트 ■ M 대 N 엘리먼트 <ul style="list-style-type: none"> ▶ 엘리먼트 이름 ▶ 엘리먼트 제약조건 <ul style="list-style-type: none"> -반정빈도수 ▶ 엘리먼트 구조(ComplexType) <ul style="list-style-type: none"> -자식 엘리먼트 발생순서 -자식 엘리먼트 반정빈도수 -자식 엘리먼트 타입 -자식 엘리먼트 노트 이름 -자식 엘리먼트 노트 구조 | <ul style="list-style-type: none"> ■ 에트리뷰트 대 에트리뷰트 ■ 1 대 1 에트리뷰트 <ul style="list-style-type: none"> ▶ 에트리뷰트 이름 -동음이의 -이음이의 ▶ 에트리뷰트 제약조건 <ul style="list-style-type: none"> -반정빈도수(occurrence) -기본값 -데이터 타입 -패킷 ▶ 에트리뷰트 기본값 <ul style="list-style-type: none"> ■ M 대 N 에트리뷰트 ■ 엘리먼트 대 에트리뷰트 <ul style="list-style-type: none"> ■ 동일한 의미의 데이터에 대한 상이한 표현 ■ 같은 정보의 다중 표현 ■ 다른 단위 ■ 다른 측정 레벨 |
|---|---|

2.2 충돌 해결 방안

스키마 충돌에 해당하는 이름 충돌은 충돌을 발생시키는 엘리먼트나 에트리뷰트의 이름을 변경시키는 방법으로 해결하고, 구조적 충돌은 엘리먼트나 에트리뷰트의 구조를 변경하여 두 스키마 사이의 구조를 동일하게 함으로써 해결한다. 또한, 데이터 타입 충돌은 충돌을 발생시키는 데이터 타입들 중에서 한쪽 데이터 타입을 다른 한쪽과 동일한 타입으로 타입 변환(type casting)하거나, 양쪽 데이터 타입 모두를 동일한 상위 데이터 타입으로 변환시킴으로써 해결한다. 마지막으로, 제약 조건 충돌은 발생 빈도/기본값 등의 제약 조건을 타당한 방향으로 일치시킴으로써 해결한다.

의미적 충돌은 XQuerySD에서의 함수 정의를 통해 의미적인 일치나 대응 관계를 기술함으로써 해결한다. 의미적 충돌 중 표현 충돌은 동일한 의미를 가지는 데이터들에 대한 대응 관계를 인위적으로 정하고, 나머지 충돌들은 수식 연산 등을 지원하는 함수를 정의함으로써 해결한다.

3. 충돌 해결을 위한 통합 스키마 저작 도구

3장에서 살펴 본 충돌에 대한 해결 방안을 참조하여 이질 정보 자원에 대한 통합을 수행할 경우, 통합 스키마를 생성해 내기 위해서는 사용자가 XQuerySD와 XML Schema의 문법을 이해하고 XQuerySD를 작성할 수 있어야 한다. 그리고 정보 자원들 간의 대응 관계를 사용자가 직접 XQuerySD 형태의 텍스트

기반으로 기술할 경우, 효율성이 떨어지고 정확성을 보장할 수 없게 된다. 이러한 문제점을 해결하기 위해, 저작자가 통합 스키마를 쉽게 생성할 수 있도록 도와주는 저작 도구(Authoring Toolkit)의 개발이 요구된다.

이 장에서는 GUI 기반의 통합 스키마 저작 도구의 설계를 위해 저작 도구가 가져야 할 기능과 규격에 대해 설명한다.

3.1 관련 연구 - XSLWiz

Induslogic 사의 XSLWiz는 드래그-앤-드롭 방식의 GUI 환경을 통해, 하나의 스키마를 따르는 XML 문서를 다른 스키마를 따르는 XML 문서로 변환해 주는 XSL 스크립트를 생성해주는 저작 도구이다[3].

XSLWiz는 XML Mapper, XML Preview, TestXML 이라는 세 개의 탭을 가지는데, XML Mapper는 소스 파일과 목적 파일을 로드시켜 트리 형태로 보여줌으로써 엘리먼트 또는 에트리뷰트들을 드래그-앤-드롭 방식으로 매핑시킬 수 있도록 해주고, XML Preview는 매핑 결과 생성된 XSL 스크립트를 보여준다. 그리고, TestXML은 매핑 전과 매핑 후의 소스 파일을 각각 보여준다. 또한, XSLWiz는 사용자가 엘리먼트 또는 에트리뷰트 간의 매핑에 필요한 산술, 문자열, 논리 연산 등을 쉽게 할 수 있도록 Funclet Wizard라는 기능을 제공한다.

그러나, XSLWiz에서는 스키마 간의 일대일 매핑만을 허용하고, XSLWiz에 미리 정의된 함수만을 매핑에 이용할 수 있다.

3.2 통합 스키마 저작 도구의 기능

본 논문에서 제안하는 스키마 통합을 지원하는 개발 도구가 가져야 할 주요 기능으로는 다음과 같은 것들이 있다.

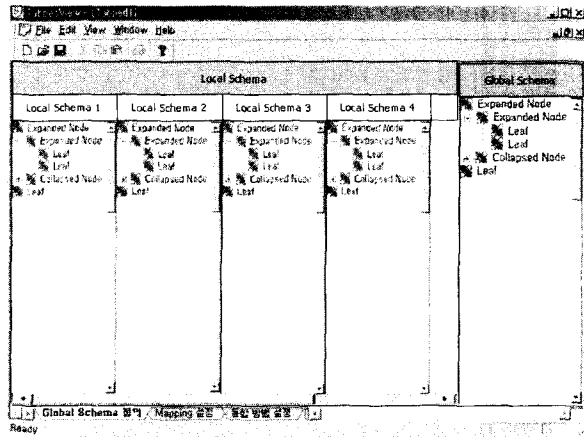
- 지역 스키마(Local Schema)의 로드 : XML Schema 파일로 된 지역 스키마를 로드하여 트리 형태로 보여준다.
- 통합 스키마(Global Schema)의 정의 : 사용자는 트리 형태로 된 지역 스키마들을 보고 통합 스키마를 트리 형태로 직접 정의한다.
- 지역 스키마와 통합 스키마의 노트 간 중재 규칙 정의 : 통합 시 연관되는 지역 스키마와 통합 스키마의 노트간 매핑을 각 노트간 연결선을 그려줌으로써 쉽게 설정할 수 있다.
- 다양한 통합 방법의 설정 : 각 스키마들간의 다양한 통합방법(UNION, JOIN, INTERSECTION)을 제공한다.
- 사용자 정의 함수의 정의 : 의미적 충돌을 해결하기 위한 기본적인 함수 라이브러리의 제공과 함께, 기본 함수를 이용하여 사용자 정의 함수를 작성할 수 있도록 한다.
- 작성된 통합 스키마 및 중재 규칙에 대한 정확성 보장

본 논문에서 제안하는 스키마 통합 저작 도구는 XQuerySD를 통한 통합 스키마 정의를 도와준다. 저작 도구에서 사용할 지역 스키마는 통합 스키마 설계 이전에 미디어이터 시스템을 통하여 얻어지고, 사용자는 지역 스키마들을 참조하여 통합 스키마를 설계한다. 통합 스키마를 설계할 때, 통합 방법을 설정하고 함수를 정의하며, 이를 근거로 XQuerySD 형태의 통합 스키마 선언이 생성된다. 그런 다음, 미디어이터 시스템은 통합 스키마에 대한 XQuerySD 선언과 보유하고 있는 지역 스키마들을 이용하여 XML Schema 형태의 통합 스키마 선언을 생성하고, 이를 통합 XML 데이터에 대한 스키마 정의로써 사용할 수 있도록 한다.

3.3 사용자 인터페이스의 구성

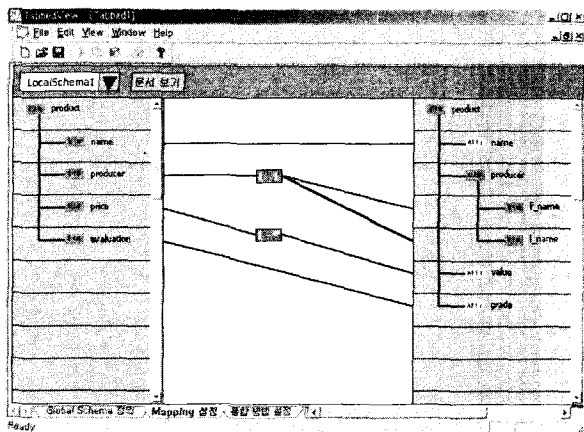
사용자 인터페이스는 3개의 작업 화면(통합 스키마 정의, 매핑 설정, 통합 방법 설정)으로 구성되고, 각 작업 화면간의 전환은 인터페이스 하단에 있는 탭을 통해 이루어진다.

통합 스키마 정의 화면은 [그림 2]에서와 같이 각 지역 스키마들을 로드해서 트리 형태로 보여주고, 로드된 지역 스키마들을 참조하여 사용자가 직접 통합 스키마를 트리 형태로 구성할 수 있다. 트리 형태로 구성된 통합 스키마는 내부적으로 XQuerySD의 return 절로 변환된다.



[그림 2] 통합 스키마 정의 화면

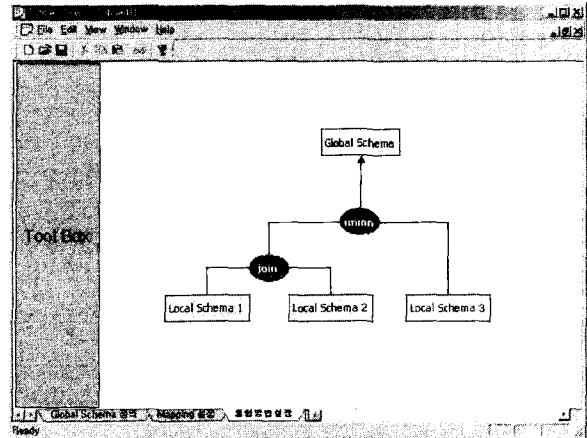
통합 스키마를 정의한 후, 매핑 설정 화면에서 각 지역 스키마와 통합 스키마 간의 노드들을 [그림 3]과 같이 매핑시킨다. 매핑 설정 화면에서 지역 스키마는 왼쪽, 통합 스키마는 오른쪽에 각각 트리 형태로 보여지고, 매핑시킬 지역 스키마는 왼쪽 상단의 리스트 박스에서 선택할 수 있다. 한번에 하나의 지역 스키마만을 보여줌으로써, 매핑을 설정할 때 화면의 복잡성을 줄이고 사용자가 매핑 설정 작업에 좀더 집중할 수 있도록 한다. 지역 스키마와 통합 스키마간의 매핑 설정은 각 노드들을 매핑시키는 연결선을 그려줌으로써 이루어지고, 사용자 정의 함수가 필요한 경우 사용자 정의 함수를 연결선상에 정의해준다.



[그림 3] 매핑 설정 화면

위의 두 작업을 마치면, 통합 방법 설정 화면에서 각 지역 스키마들을 어떤 방법(UNION, INTERSECTION, JOIN)으로

통합할 것인지를 결정해야 한다. [그림 4]에서 보는 바와 같이 통합 방법 설정 화면의 왼쪽에 도구 상자가 존재하는데, 이 도구 상자는 통합 방법 설정 시 필요한 도구들을 제공한다. 예를 들어, 지역 스키마들을 JOIN으로 통합하기 위한 연결선 같은 도구들을 제공한다. 통합 방법 설정 화면을 분리시킴으로써, 좀더 복잡한 통합 방법의 설정을 가능하게 한다.



[그림 4] 통합 방법 설정 화면

4. 결론 및 향후 과제

XML Schema는 기존의 DTD에 비해 많은 데이터 타입과 제약 사항, 다양한 구조의 엘리먼트와 애트리뷰트 선언을 지원함으로써 보다 풍부한 XML 인스턴스에 대한 스키마 정의가 가능하게 하였지만, XML Schema의 이러한 기능들은 기존의 DTD를 스키마 정보로 이용하는 XML 인스턴스보다 많은 충돌 문제를 발생시킨다.

따라서, 본 논문에서는 XML을 공통 데이터 모델로 이용하고, XML에 대한 스키마 정의 언어로 XML Schema를 이용하여 다양한 정보들을 통합할 경우, XML Schema에서 발생하는 충돌 문제를 분류하고 각 충돌 문제에 대한 해결 방안을 제시함과 동시에 수행한 충돌 분류와 그 해결 방안을 바탕으로 통합 스키마 작성을 용이하게 할 수 있도록 지원해 주는 저작 도구의 기능과 구조에 대한 연구를 수행하였다. 그리고, 향후 연구 과제로는 본 논문에서 설계한 저작 도구의 구현과 구현한 저작 도구를 적용시스템인 DataBlender에 결합시키는 문제가 남아 있다.

[참고 문헌]

- [1] 권석훈, 이승원, 박재홍, 김미혜, 이정하, 이규철, "이질 자료 모델의 충돌 해결 방안 연구" 최종 보고서, 충남대학교, 2002.1
- [2] 이승원, 권석훈, 김미혜, 이정하, 이규철, "XML Schema를 이용한 스키마 통합시 충돌 문제의 분류", 한국정보과학회 학술발표논문집, 제28권 2호, pp31-33, 2001.10
- [3] XSLWiz™ 2.1, Induslogic Inc, <http://www.induslogic.com/products/products2.html>
- [4] W3C, Recommendation, "XML Schema Part 0 : Primer", <http://www.w3.org/TR/xmlschema-0/>