XML Schema기반 시맨틱 데이터 통합

김동광°, 정갑주 신효섭, 황선태
건국대학교 신기술융합학과, 인터넷&멀티미디어공학과, 국민대학교 컴퓨터공학과
walhalla@gcslab.konkuk.ac.kr°, {jeongk, hsshin}@konkuk.ac.kr, sthwang@kookmin.ac.kr

An XML Schema-based Semantic Data Integration

Dongkwang Kim°, Karpjoo Jeong Hyoseop Shin, Suntae Hwang
Department of Internet & Multimedia Engineering Konkuk Univ,
Department of Computer Science Kookmin Univ

요 약

최근 들어서 생물학, 건설, 토목, 의학, 화학, 항공 등 다양한 과학 분야에서 효율적인 공동 연구를 돕기 위해서 그리드 기반 데이터 통합을 위한 많은 연구가 진행되고 있다. 하지만 각 분야에서 사용하는 이질적인 형식의 데이터를 원활히 통합하는 것은 매우 어려운 문제이다. 이러한 측면에서, XML은 이질적인 시스템 간 데이터 공유의 수단으로서 광범위하게 사용되고 있으나, 데이터를 의미적으로 통합하는데 에는 한계가 있다. 이 문제를 해결하기 위해서, 본 논문에서는 그리드 환경에서 XML Schema 에 기반을 둔 이질적인 데이터 소스들을 의미적으로 통합하고 검색하는 방법으로서 RDF/RDFS 기반의 온톨리지 매핑을 통한 통합 검색 방안을 제안한다. 우리가 제안하는 방법은 특히 데이터를 XML Schema를 이용해서 표현하기 때문에 원래 데이터소스에 대한 수정 없이도, XML Schema 기반의 구조적 검색뿐 아니라, 온톨로지 기반의 의미적인 데이터 통합 및 검색을 가능하게 한다. 우리는 이 방법을 토목 분야의 시스템에 적용하여 검증하였다.

1. 서 론

다양한 과학 분야의 연구원들이 연구의 효율성을 극대화하기 위해서 같은 연구 분야뿐만 아니라 다른 분야들 간 공동 연구를 위한 노력이 계속되고 있으나, 이런 공동 연구는 지역적, 공간적인 제약 때문에 많은 어려움이 있었다. 그러나 최근 들어 통합·공유 환경 구축을 위한 인프라 기술인 그리드 기술을 이용해서 공동 연구가 활발하게 이루어지고 있다. 하지만 다양한 과학 분야 간의 통합은 이질적인 데이터들 때문에 어렵다. 각 연구 기관에서는 연구의 주제 및 연구 조건과 연구에 사용하는 장비들에 따라서 다양한 데이터(파일, 이미지, 동영상, 데이터베이스 테이블, 등)들이 생성되고, 사용되어지고 있다. 이러한 이질적인 형식의 데이터들을 원활히 공유하기 위해서 XML이 광범위하게 사용되고 있다. XML은 데이터의 종류에 독립적으로 표준화된 방식으로 데이터를 표현 할 수 있기 때문에 그리드와 같은 이질적인 환경에서 데이터를 공유하기에는 적합하다. 하지만, 계층적인 특징을 가지고 있는 XML은 구조적인 차이에 의해서 통합하기 어렵고 연구 조건의 변화에 유연하게 대처하기가 어려울 뿐만 아니라, 의미적으로 통합하는 데에도 한계가 있다.

이러한 문제를 해결하기 위해서, 본 논문에서는 이질적인 데이터들을 메타 데이터를 이용해서 표현하고, 이런 메타데이터의 데이터 모델을 유연하고, 확장성 있는 XML Schema로 정의해서 데이터 모델을 통합 관리하여, 구조적인 검색이 가능하게 할 것이다. 또한 구조적인 차이로 인한 통합 검색의 한계를, RDF/RDF Schema를 이용한 통합 온톨로지 구축을 통해서 통합 검색뿐만 아니라, 의미적인 통합 방법을 제안할 것이다.

2. 관련 연구

데이터 통합을 위한 연구는 많은 분야에서 활발하게 이루어지고 있다. 그리드 분야의 NEESgrid 프로젝트는 데이터 통합을 위해서 참조 데이터 모델을 정의하여 각 연구기관들의 데이터를 통합을 시도하였다[1,2]. 또한 이 데이터 모델을 기반으로 중앙 데이터 저장소를 구축하여 연구원들 간의 공동 연구 및 데이터 공유를 위한 서비스를 제공한다. 그러나 데이터 모델이 고정되어 있기 때문에 다양한 연구 환경을 포괄적으로 저장하는데 에는 한계가 있다. GEONgrid(Geoscience Network)[3]는 지구과학 분야의 공동 연구를 위한 이미지 데이터를 통합 관리하기 위한 시스템으로서, 이 이미지들의 메타데이터를 시맨틱하게 표현해서 관리하고, 통합 검색할 수 있도록 구축하였다. 그러나 이 프로젝트는 데이터를 통합한 것이 아니라, 지구 과학의 분류 기준으로 사용되는 정보들을 온톨로지로 구축하고, 연구를 통해서 얻어지는 이미지 데이터들과 온톨로지로 표현된 분류 정보들의 관계를 온톨로지로 표현해서 관리하고, 분류 정보를 검색조건으로 이용한 검색이 가능하도록 구축되어있다.

EDUTELLA[4]은 시맨틱한 데이터 공유방법을 P2P Network을 에 적용한 시스템이다. 이 프로젝트는 RDF를 이용해서 실제 데이터를 표현하고, 검색을 위해서 RDF질의어를 표현하고, 이기종간의 데이터들에 대해서 검색 할 수 있는 변환 모듈을 개발하였다. 그러나 EDUTELLA는 다른 시스템간의 맵핑을 위해서 후보들을 선정해서 맵핑하는 방법을 채택하여 광범위한 데이터를 통합 검색하기에는 한계를 가지고 있다. Piazza[5]는 XML로 표현되어 있는 데이터들을 공유하기 위해서 XML간의 구조적인 차이점을 설명하는 XQuery와 유사한 언어를 개발하였다. 이 언어를 이용해서 각 XML문서의 차이점을 기술하고, 자동 번역 시스템을 통해서 각 시스템간의 데이

터를 공유하도록 구축되어 있다.

3. 시맨틱 데이터 통합

다양한 분야에서 연구의 결과 데이터뿐만 아니라 연구 수행 중에도 다양한 데이터(실험 조건, 연구원 정보, 실험 장비 등)들이 생성되고, 또한 실험 결과를 분석해서 작성하는 보고서나 논문과 같은 문서들도 만들어진다. 이런 데이터들을 통합 관리하기 위해 각 데이터를 포함하는 메타 데이터를 정의하고, 메타데이터의 데이터 모델을 유연하고, 확장성이 뛰어난 XML Schema로 정의하였다. 하지만 이렇게 정의된 데이터 모델을 통합하기 위한 차이점을 다음과 같은 3가지 구조적·의미적인 요소로 정리할 수 있다.<그림1>

1. 이름은 다르지만 의미가 같은 것: Schema A에서 정의한 Element와 Schema B에서 정의되어 있는 Element가 정의된 이름은 다르지만, 의미하는 것은 같은 경우 (/A/B 와 /A/B' 은 같다).
2. 이름은 같지만 의미가 다른 것: Schema A에 정의한 Element와 Schema B에 정의되어 있는 Element가 이름이 같지만, 이들이 의미하는 것은 틀릴 경우 (두 Schema의 /A/C는 다르다).
3. 이름과 의미가 같은 경우: SchemaA의 element와 Schema B의 Element가 구조적으로 같고, 이름과 의미가 같은 경우 (두 Schema의 /A/D는 같다)

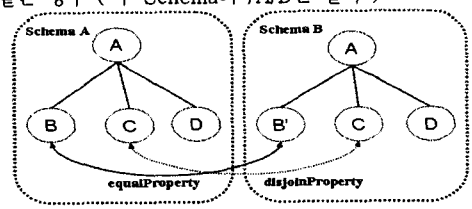

그림 1 데이터 모델간의 구조적·의미적 차이

이러한 구조적·의미적 차이점을 통합하고 의미적인 관리 및 통합 검색을 하기 위해서 메타데이터의 요소들을 RDF[6]로 표현할 수 있는 변환 규칙을 정의 했다. 먼저 XML Schema의 각 Element들을 *subject*로, Element가 가지고 있는 Sub-Element들을 *object*로 정하고 Element와 Sub-Element들의 관계를 나타내는 *predicate*를 정의했다. XML Schema를 RDF로 표현하기 위해서 우리는 3개의 *predicate*를 정의하였다.

1. *hasProperty*: subject와 object 관계인 Element와 Sub-Element간의 관계를 나타내는 *predicate*
2. *equalProperty*: 두 데이터 모델 사이에서 이름은 다르지만 의미가 같은 Element들의 관계를 나타내는 *predicate*
3. *disjoinProperty*: 두 데이터 모델 사이에서 이름은 같지만 의미가 다른 Element들의 관계를 나타내는 *predicate*

<그림 2>는 위에서 정의한 방법에 이용해서 서로 다른 두 개의 데이터 모델들을 통합하기 위한 온톨로지를 구축하는 것을 설명하고 있다. XML Schema의 각 Element들은 URI로 정의해서 통합 온톨로지에 추가되고, 상호간의 관계가 설정된다. 하지만, 이름, 의미, 구조측면에서 동일한 Element인 Equipments와 DAQBoard는 한번만 추가된다.

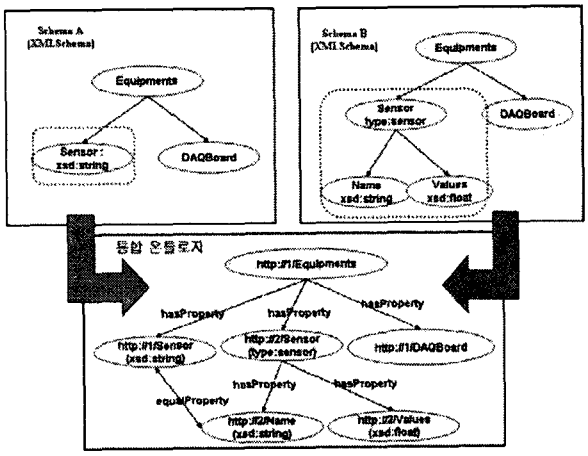

그림 2 통합 온톨로지 예제

통합 검색을 위해서 구축되어 있는 통합 온톨로지를 이용해서 사용자는 검색 조건을 만들게 된다. 통한 온톨로지에서 검색하고자 하는 Element를 선택하고 검색 어를 입력한다. 그 다음에는 선택한 Element와 *predicate*가 *equalProperty/disjoinProperty*관계인 Element들을 검색하기 위해서 RDQL[7]을 생성한다. 생성된 RDQL을 순차적으로 실행해서 *equalProperty/disjoinProperty*관계가 있는 Element들을 찾아낸다. 검색된 Element들은 자신과 *predicate*가 *hasProperty*관계인 subject를 찾아서 구조적인 관계를 파악하고, 이 정보를 기반으로 XQuery를 생성한다. 이렇게 생성된 XQuery들은 해당 Schema가 저장되어 있는 데이터베이스에서 실행되고, 결과 데이터는 통합되어 사용자에게 보이게 된다.

<그림 3>은 우리 시스템의 전체적인 구조를 보여주고 이 시스템은 크게 두 단계로 나눌 수 있다. 데이터 모델을 정의하고 온톨로지를 구축하는: '1) 스키마 생성'과, 통합 온톨로지를 이용한 '2) 데이터 검색' 이다.

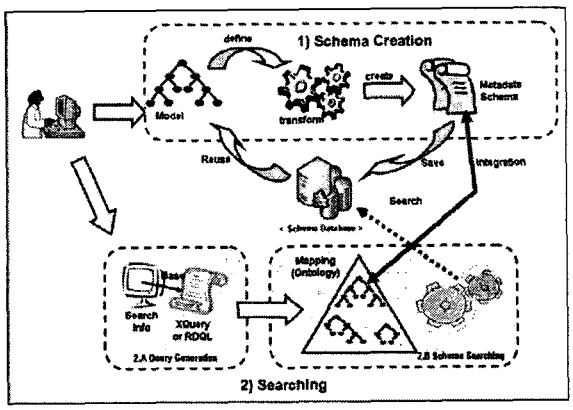

그림 3 전체 시스템 구조

4. 사용자 인터페이스 및 구현

우리는 건축토목분야의 그리드 시스템인 KOCED[8]에 본 논문의 방법을 적용하였다. KOCED의 다양한 건축·토목의 연구 분야에서 실험 정보를 공유하기 위해 각 실험에 맞는 데이터 모델을 정의하고 관리하기 위한 XML Schema기반의 Tool을 구현하였다. 이 Tool을 이용해서 새로운 데이터 모델을 XML Schema문법에 대해서 알고 있지 않아도 쉽게 작성할 수 있고, 또한 기존에 등록된 Schema를 검색해서 자신이 표현하고자 하는 데이터 모델로 수정하거나, 새로운 데이터 모델로 등록 할 수 있게 구현하였다. 시스템에 등록되어 있는 메타 데이터의 데이터 모델을 가지고 실제 데이터를 입력할 수 있도록, 데이터 모델 기반의 입력폼을 자동으로 생성하고, 입력할 수 있는 기능을 제공하며, 이렇게 입력된 데이터는 XML Database에 저장되고 관리된다. 또한 새로운 데이터 모델을 추가할 때, 현재까지 구축된 통합 Ontology에 Mapping하기 위한 Tool을 제공하고 있다. 이를 이용해서 사용자는 간편하게 새롭게 추가된 Element와 기존의 통합 Ontology와 Mapping 작업을 처리할 수 있다.

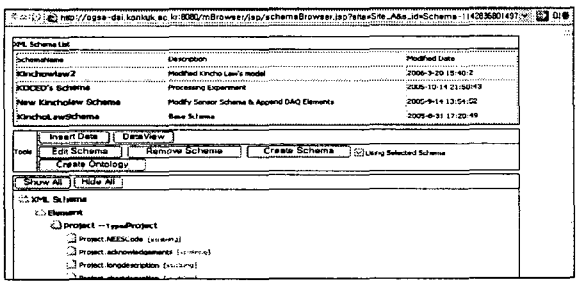

그림 4 XML Schema기반 데이터 모델 관리

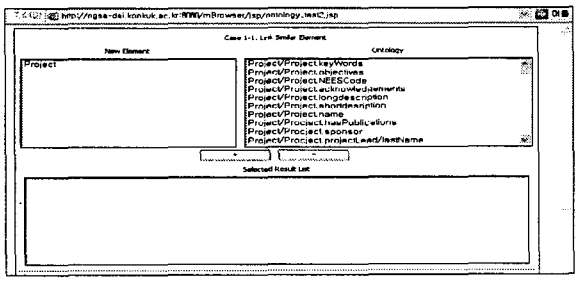

그림 5 통합 온톨로지 맵핑

또한, 사용자들이 데이터 모델에 대한 정보 없이 시스템에 등록되어 있는 메타데이터를 검색할 수 있도록, 구축된 Ontology를 검색 조건으로 보여주고, 이 조건 중에 하나를 선택해서 질의어를 작성 할 수 있다. 그리고 하나의 조건만이 아닌 복수의 질의어를 입력하여, 효율적인 검색이 가능하도록 구현하였다. 이렇게 생성된 검색 질의어는 OGSA-DAI[9]를 통해서 그리드 시스템의 전체 데이터베이스에 통합 질의를 실행하고 결과를 취합해서 사용자에게 보여준다.

5. 결론 및 향후 연구 과제

그리드 시스템에서의 데이터 통합은 과학 분야의 효율적인 공동연구를 위해서 꼭 해결돼야 하는 중요한 부분이다. 다양한 형태의 데이터들을 쉽게 표현하고, 연구 기관들 간에 효율적으로 접근하기 위해서는 통합 검색이 필요하다.

이 논문에서는 이질적인 데이터에 대한 효율적 공유를 위한 XML Schema기반의 데이터 통합 방법을 제안하였다. 먼저 같은 연구 분야의 데이터들이 실험 시설이나 실험 조건에 따라서 다양하게 정의되야하는 특징이 있는 메타데이터를 확장성(flexible)있는 XML Schema를 이용해 정의하였다. 또한, 사용자가 다양한 데이터 모델을 통합 검색하기 위해서 시스템에 등록되는 메타데이터의 데이터 모델의 각 Element간의 관계만을 통합 온톨로지로 구축하므로서, 계속 변화하는 데이터 모델에 대한 효율적인 관리와, 검색 방안을 개발하였다.

우리가 제안한 방안인 XML Schema로 정의되어 있는 데이터 모델과 RDF로 표현되는 Ontology와의 Mapping과 통합 검색 시에 선택하는 Ontology에 대해서, 사용자가 쉽게 접근 할 수 있는 Visual Tool의 개발에 대해서 계속적인 연구가 필요하다.

6. 후기

본 결과물은 건설교통부에서 수행하는 KOCED의 연구 결과입니다.

7. 참고문헌

1. NEESgrid , http://it.nees.org
2. Reference NEESgrid Data Model [TR-2004-40] (2004), Jun Peng , Kincho H. Law,
3. GEONgrid, http://www.geongrid.org
4. EDUTELLA: A P2P Networking Infrastructure Based on RDF (2002), Wolfgang Nejdl, Boris Wolf, Changtao Qu, May 7-11, 2002, WWW2002
5. Piazza: Mediation and Integration Infrastructure for Semantic Web Data, Zachary G. Ives, Alon Y. Halevy, Peter Mork, Journal of Web Semantics manuscript.
6. RDF(Resource Description Framework), http://www.w3.org/RDF/
7. RDQL, Andy Seaborne, HP Labs Bristol (2004), http://www.w3.org/Submission/RDQL/
8. Korea Construction Engineering Development Collaboration (KOCED) , http://www.koced.net
9.OGSA-DAI(Database Access and Integration Services) , http://www.ogsadai.org.uk