XML 문서선별과 질의확장을 위한 자동화 모듈 개발

김명숙, 권혁돈, 공용해
순천향대학교 정보기술공학부

XML Document Selection and Query Expansion Modules

Myoung Sook Kim, Hyek Don Kwon, Yong Hae Kong
Division of Information Technology Engineering Soonchunhyang Univ.

요 약

본 연구는 다양한 형식을 가지는 XML 문서의 효율적인 정보검색을 위한 다음과 같은 자동화 모듈들을 개발하였다. 구현된 모듈은 XML 문서를 획득하는 문서추출 모듈, 온톨로지를 이용한 포괄적 DTD 생성 모듈, 생성된 포괄적 DTD와 XML 파서를 이용하여 정보검색 대상 XML 문서를 사전에 선별하는 문서여과 모듈, XML 질의를 확장하는 질의확장 모듈, JDOM의 XPath를 이용한 질의엔진 모듈로 구성된다. 이와 같이 구현한 모듈들을 샘플 XML 문서에 적용하여 XML 문서추출, DTD 생성, 문서여과, 질의확장, 질의엔진의 효과를 실험하였다.

1. 서론

다양한 구조와 속성을 지닌 XML 문서를 대상으로 효과적인 질의를 위해 문서여과를 위한 포괄적 DTD 생성 알고리즘과 질의를 의미적으로 확장하는 알고리즘을 개발한 바 있다[1]. 포괄적 DTD는 특정 영역에 대한 온톨로지를 근간으로 하여 구조가 다른 XML 문서에 적용 가능하게 설계되었으며, 이 포괄적 DTD를 바탕으로 불필요한 문서들을 여과할 수 있다. 또한 질의확장 알고리즘은 온톨로지의 개념구조와 상호 연관관계를 추론하여 질의를 확장함으로써 보다 풍부하고 효과적인 의미정보 검색을 가능하도록 하였다.

본 연구는 검색대상 문서의 여과와 질의확장을 통해 의미정보를 효과적으로 검색할 수 있는 다음의 모듈들을 구현하였다. 먼저 URL 리스트 파일로부터 해당 XML 문서들을 추출하는 문서추출기를 구현하였다. 다음으로 추출한 문서들 중 관심 대상이 아닌 문서를 여과하기 위하여 포괄적 DTD 생성기를 구현하였다. 문서여과 모듈은 포괄적 DTD와 XML 파서에 의해 관심의 대상이 아닌 문서를 사전에 여과한다. 이렇게 여과된 문서만이 질의 검색대상이 되고, 이때 질의방식은 XPath 규칙을 따른다. 질의확장 모듈은 인가된 질의를 계층적 구조분석과 연관관계를 추론한 규칙들을 참조하여 확장한다. 질의엔진은 XML 문서의 구조적 검색을 통하여 해당 엘리먼트를 검색한 질의 결과를 보여준다. 마지막으로 구현한 모듈을 샘플 XML 문서를 이용하여 테스트하였다.

2. 문서추출기 (XML Spider) 개발

XML Spider는 웹상에 존재하는 해당 XML 문서를 추출한다. XML 문서의 위치는 URL 리스트 파일로써 입력을 받으며, URL 리스트 파일에는 XML 문서 추출을 위한 URL이 입력 되어있다. 그림 1은 웹에서 XML Spider를 이용해 XML 문서를 추출하는 과정이다.

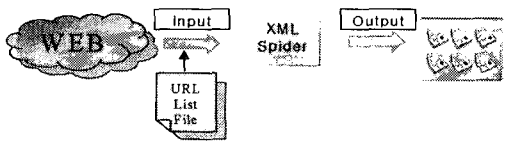

그림 1. XML Spider

문서 추출기에 URL이 입력되면 해당 URL에 존재

본 논문은 정보통신부 정보통신연구진흥원에서 지원하고 있는 정보통신기초기술연구지원사업의 연구결과입니다.

하는 XML 문서를 파일로 추출하는 문서추출기의 문서여과 과정은 다음과 같다. 먼저 웹상에 존재하는 XML 문서를 URL 리스트 파일로 입력받아, 해당 URL에서 XML 문서를 검색하여 호스트 컴퓨터에 파일 단위로 저장한다. 이 과정에서 XML Spider는 java.net package를 이용한다. 본 논문에서 적용한 샘플 XML 문서는 실험 URL에 해당하는 웹상에 존재하는 문서이고, 대학연구센터 온톨로지를 대상으로 실험하였다.

다음은 XML Spider의 실행결과이다. 추출하려는 URL 리스트 파일의 내용은 그림 2와 같다.

그림 2. URL 리스트 파일

그림 3-1, 3-2는 웹상에 존재하는 샘플 XML 문서 example01.xml, example02.xml 이다.

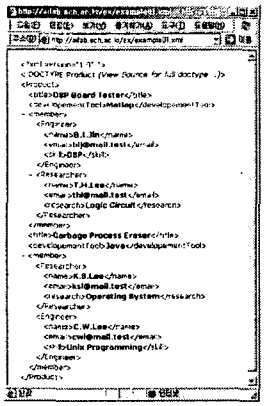

그림 3-1.
http://ailab.sch.ac.kr/ex/
example01.xml

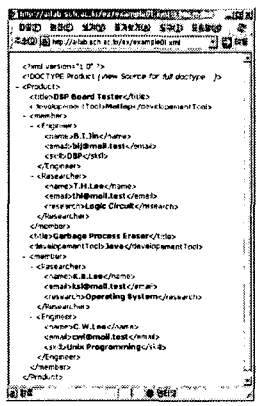

그림 3-2.
http://ailab.sch.ac.kr/ex/
example02.xml

문서추출기의 문서여과 실험을 위해, 샘플 XML 문서 example02.xml은 대학연구센터 온톨로지를 기반으로 작성하였으며, example01.xml은 example02.xml과 동일한 엘리먼트를 사용하였지만 온톨로지에서 정의한 구조와는 다르게 작성하였다. 그림 4-1과 그림 4-2는 문서추출기를 이용하여 URL로부터 추출한 XML 문서이다.

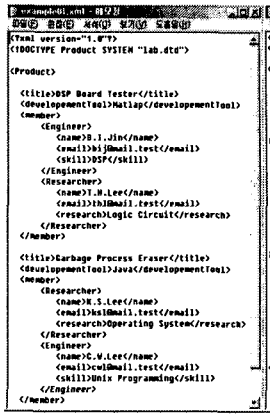

그림 4-1.
example01.xml

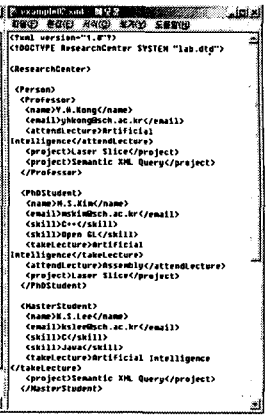

그림 4-2.
example02.xml

3. 포괄적 DTD 생성기 개발

포괄적 DTD 생성기는 특정영역 온톨로지의 계층적 구조와 연관관계에 의해서 온톨로지 기반의 XML 문서에 유효한 DTD를 생성한다. DTD 생성기는 온톨로지 프로세서와 DTD 생성 프로세서로 구성된다. 온톨로지 프로세서는 온톨로지의 개념과 속성을 파싱하여 개념 및 속성 클래스를 생성하고, DTD 생성 프로세서는 생성된 개념 및 속성 클래스의 ENTITY 생성, ELEMENT 생성, ATTLIST 생성, 추가 ELEMENT 생성 알고리즘을 각각 적용하여 포괄적 DTD를 생성한다[1].

4. 문서여과 모듈 개발

추출한 XML 문서여과 모듈은 XML 파서와 포괄적 DTD를 이용한다. XML 파서는 XML 문서가 XML의 구조적 제한조건(Constraint)을 따르는지 검사한다[2].

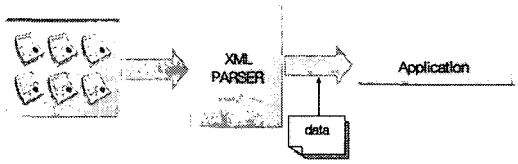

그림 5. XML parser

또한 XML 파서는 그림 5에서와 같이 Application에게 XML 문서의 data 또는 자료구조를 제공한다. 본 연구에서는 문서 전체를 읽지 않고 XML 문서를 순차적으로 읽는 SAX 파서를 사용한다. SAX는 XML 문서에 접근하여 필요한 데이터를 가져오거나

수정하는 API이다.

Validating 파서는 XML 문서가 Well-formed 문서인지와 DTD 구조에 적합한지를 검증한다. 따라서 포괄적 DTD에 의한 문서의 유효성 판단을 위해 validating 파서인 SAX를 사용하고 XML 문서를 여과한다. XML 문서는 특정한 데이터의 구조적 집합체이므로 SAX 파서의 사용은 필수적이다. 또한 SAX는 문서의 파싱 과정에서 계속 이벤트를 발생시키기 때문에 문서 전체를 읽지 않고도 XML 문서의 유효성 검증이 가능하다.

그림 4-1의 example01.xml은 그림 6과 같이 유효성 검사과정에서 여과된다. 결과적으로 문서의 여과 과정에서 example02.xml 문서만이 유효한 문서가 되었다.

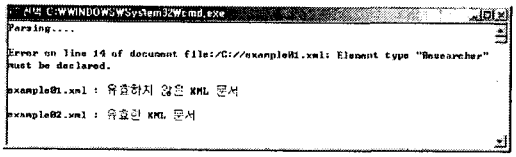

그림 6. 문서여과 결과

5. 질의확장 모듈 개발

질의확장은 온톨로지의 개념 계층구조와 연관관계를 이용한다. 개념 계층구조는 온톨로지에 표면적으로 나타나 있지만, 연관관계는 묵시적으로 표현되어 있기 때문에 규칙 추출기를 이용하여 질의를 확장한다.

그림 7-1은 대학연구센터의 온톨로지의 계층적 구조를 나타낸다. 대학연구센터 온톨로지의 연관관계는 그림 7-2와 같다.

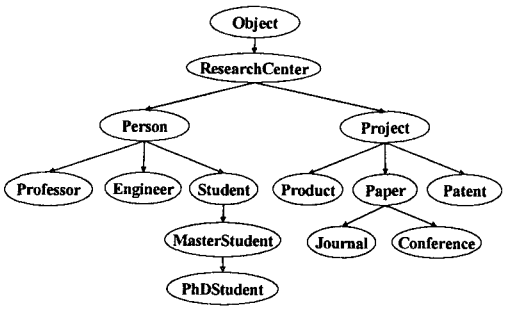

그림 7-1. 대학연구센터 온톨로지의 계층적 구조

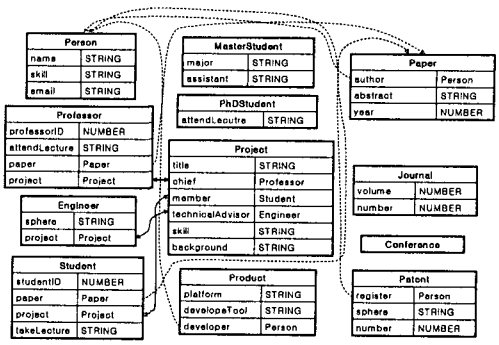

그림 7-2. 대학연구센터 온톨로지 연관관계

그림 8은 온톨로지를 이용한 XML 질의확장 과정이다.

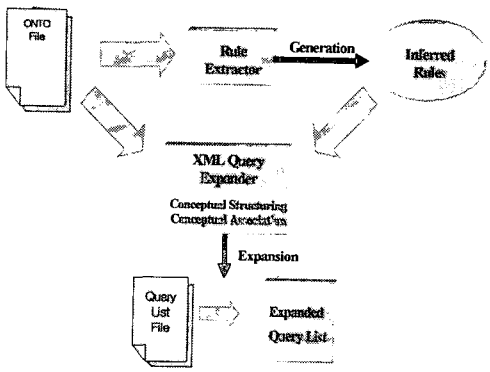

그림 8. 온톨로지를 이용한 XML 질의확장

사용자가 입력한 하나의 질의는 질의확장기에 의해 질의리스트로 다수의 확장질의가 반환된다. 다음으로 질의엔진은 질의리스트를 입력받아 XML 문서를 구조적으로 검색하여 유용한 정보를 획득한다.

6. 질의엔진 (Query Engine) 개발

XML 질의확장 시스템에서 SAX 파서에 의해 여과된 XML 문서에 대해 확장된 XML 질의를 입력하여 의미정보를 추출하게 된다. 이때 XML 문서의 구조적 검색을 위하여 JDOM의 XPath를 이용한다[3]. XPath는 XML 문서의 부분을 정의하기 위한 문법 규약(syntax rule)의 집합이라고 할 수 있으며 XPath는 XML 문서의 노드를 정의하기 위하여 경로식(path expression)을 사용한다. XPath의 중요 목적은 XML 문서의 부분들을 노드의 트리구조를 이용하여 찾아갈 수 있도록 논리적인 XML 문서 모델을 만드는 것이

다. XPath 패턴은 '/'로 분리된 자식 엘리먼트의 이름 리스트로써 XML 문서에서의 경로를 표현하며, 이러한 패턴의 형식으로 경로를 탐색하여 엘리먼트를 검색한다. XML 문서를 검색한 결과는 검색된 엘리먼트 리스트로서 출력 하게 된다[3],[4].

질의엔진은 질의확장 모듈에서 확장된 질의리스트로부터 해당 XML 문서에 질의를 하게 된다. 그림 9는 여과된 example01.xml 문서에 'C'에 관한 skill을 지닌 사람을 검색하기 위한 //Person[skill="C"]라는 질의의 검색 결과이다.

질의확장 모듈에서 확장된 질의리스트는 //(Person | Professor | Enginner | Student | MasterStudent | PhDstudent) [skill="c"]와 같은 값을 반환하고, Query Engine은 질의리스트에 대하여 정보를 추출한다. 그 결과, MasterStudent 엘리먼트 항목에서 "C"라는 정보를 검색하게 된다.

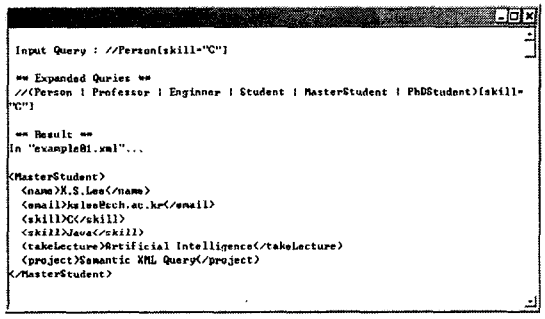

그림 9. 확장된 질의의 결과 엘리먼트 리스트

7. 결론

본 연구는 다양한 구조와 속성을 지닌 XML 문서에 내포되어 있는 정보를 효과적으로 검색하기 위하여 문서여과 모듈, 질의확장 모듈, 검색엔진 모듈 등을 구현하였다. 우선, 문서여과 모듈은 특정 웹에서 추출한 XML 문서를 포괄적 DTD에 의해 여과하고, 질의확장 모듈은 여과과정을 거친 XML 문서를 대상으로 질의를 확장한다. 다음으로 검색엔진 모듈은 확장된 질의를 사용하여 문서에 내포되어 있는 정보를 검색한다. 이러한 일련의 과정을 수행하는 모듈들을 단계별로 세분화하여 구현하였다.

세부적으로는 살펴보면, 문서추출기는 웹으로부터 XML 문서를 추출하고, 추출한 XML 문서에서 불필요한 문서를 여과하기 위하여 문서여과 모듈을 구현하였다. 포괄적 DTD 생성기는 여러 XML 문서에 적용 가능한 포괄적 DTD를 생성하고, 생성된 포괄적 DTD를 기반으로 문서여과 모듈의 XML 파서는 정보검색의 대상이 되는 XML 문서를 사전에 여과한다. 여과된 XML 문서를 대상으로 질의를 확장하고 확장 질의에 의한 의미정보 검색을 수행하는 질의확장 모듈과 질의엔진을 구현하였다. 질의확장 모듈은 온톨로지의 개념 계층 구조와 규칙 추출기에 의해 추출된 연관관계를 바탕으로 확장된 질의리스트를 반환하고, 질의엔진은 질의리스트를 입력받아 JDOM의 XPath를 이용하여 XML 문서를 구조적으로 검색한다.

XML 문서선별과 질의확장을 위한 자동화 모듈은 XML 문서검색으로부터 확장질의에 의한 결과출력에 이르기까지 모든 과정을 효과적으로 수행하였다.

[참고문헌]

[1] Kyeung Soo Lee, Dong Ik Oh, and Yong Hae Kong, "Semantic XML Filtering by Ontology Combination, Web Engineering(LNCS 2722), pp.395-398, 2003.

[2] Jason Hunter, "JDOM and XML Parsing", http://otn.oracle.com/oramag/oracle/02-sep/o52jdom.html, 2002.

[3] W3C, "XML Path Language(XPath) 2.0", http://www.w3.org/TR/xpath, 1999.

[4] Jason Hunter, Brett McLaughlin, "JDOM v1.0beta10-dev API Specification", http://www.jdom.org/docs/apidocs/index.html, 2004.