이름 공간을 이용한 질의 검색 시스템 설계 및 구현

이원철, 이상민
강원대학교 전자계산학과
woncheol@mirae.kangwon.ac.kr, smrhee@cc.kangwon.ac.kr

Design of retrieval system using Namespace

won-cheol lee, sang-min lee
Dept. of Computer Science kangwon National University

요 약

xml이 확산되면서 이를 저장하고 검색하는 방법들이 많이 제안되었다. 또한 데이터의 통합에 대하여 여러 가지 방법을 제시하고 있다. 그러나 이런 방법들은 xml이 가지고 있는 자유로운 확장성과 데이터의 통합이라는 관점에서 서로 상반된 면을 나타낸다. 이런 문제의 해결을 위해 W3C에서 제안한 이름공간에 대한 기존의 기능을 확장하고, 검색에 있어 사용자의 편리성을 위한 재질의 기법과 정확한 검색 결과를 위한 이름공간을 이용한 검색 시스템을 설계하였다.

1. 서 론

기존의 HTML(Hyper Text Markup Language)이 갖는 단점과 SGML(Standard Generation Markup Language)이 갖는 단점을 보완하여 작성된 차세대 웹 언어 표준인 XML (Extensible Markup Language)은 1996년 W3C(World Wide Web Consortium)의 XML Working Group에서 제안하였다.

이러한 XML은 웹 문서의 효과적인 관리를 위해 W3C를 주축으로한 관련 연구들이 활발히 진행되었다. XML 문서들을 저장 관리 및 검색에 관한 연구[10], XML 저장관리 시스템 개발에 관한 연구[13], 기존의 데이터베이스 시스템에 저장되어 있는 데이터를 XML 데이터로 변환 및 저장하는 연구[14] 등이 있다.

구조적으로 구성된 XML은 정보의 검색에 있어서 HTML에서 제공해 주기 못하는 새로운 검색 기법을 제공해 준다. 즉 논리적 구조를 표현하는 여러 가지 DTD나 XMLSchema를 사용함으로써 XML 문서의 저장 및 관리와 검색에 효율적으로 사용된다.

컴퓨터 시스템의 발달과 WWW(World Wide Web)기술의 보편화는 수많은 정보 시스템과 다양하고 분산된 정보 서비스로 인하여 거대한 인터넷 정보 자원을 구축하게 되었고, 이로 인하여 분산되어 있는 다양한 형태의 정보 자원을 통합하여 검색할 필요성이 제기 되었다.

이러한 요구 사항의 해결 방안으로 DTD를 데이터의 스키마 정보로 활용한 데이터 통합에 대한 연구가 진행되었으나 DTD 자체가 가지고 있는 단점들로 인하여 많은 문제점이 지적 되었고 이를 해결하기 위해 W3C에서 XMLSchema를 제안하였다.[12] 그러나 XMLSchema 역시 분산되어 있는 데이터의 검색과 통합에 있어 문제점이 발견되었다. 또한 검색시 필요한 스키마의 구조와 엘리먼트의 이름 등에 대한 정보를 알고 있어야 엘리먼트를 이용한 구조적 검색을 할 수 있으므로 사용자에게 불편을 주었다.

따라서 본 논문에서는 위의 문제점을 극소화시키고, 검색의 효율성을 높이기 위하여 사용자의 편의성 및 정확성을 제공하는 이름공간을 이용한 검색 방법을 제시한다.

2.관련연구

2-1.데이터 통합 검색시 발생하는 문제점

컴퓨터의 데이터의 저장공간이 다양해지고, WWW(World Wide Web)을 통한 데이터의 교환이 보편화되어 수많은 정보 시스템과 분산된 정보는 거대한 인터넷 데이터베이스를 만들었다. 이로 인하여 분산되어 있는 다양한 형태의 정보 자원의 통합적 저장과 검색이 어렵게 되었다.

XML 데이터의 통합에 있어 가장 활발한 연구를 이루는 것이 스키마의 통합이다. Schema란 일반적으로 데이터의 모델링을 통해 유도되는 데이터 구조 및 짜임새이다. 여기서 구조는 전형적으로 데이터 항목들에 대한 이름과 이에 적용되는 구속 조건들의 목록으로 만들어진 일종의 통제된 어휘집을 사용하여 기술된다.

스키마의 통합에 있어 발생할 수 있는 문제[4]들이 대부분 정보의 통합 검색시 발생하는 문제들이다. XML이 확장성이 용이하기 때문에 더 많은 문제들을 발생시킨다. 이러한 문제를 해결하기 위해서 W3C에서 새로운 스키마 정의로 XML스키마를 표준화하고 있지만 방대해진 DTD는 데이터의 통합에 많은 문제를 제기한다. 데이터의 검색에 있어서 발생할 수 있는 문제는 크게 두 개로 나누어 볼 수 있다. 스키마의 이질성에 의해 발생되는 충돌과 (schematic conflict)와 데이터간의 이질성에 의해 발생되는 의미 충돌(semantic conflict) 등이 있다.

2-2.xml 검색기법

질의어를 통한 검색[5]XML-QL, Quilt, Xlink, Xpath 등과 검색 및 저장을 위한 데이터 시스템 설계 [10], 의미 통

합을 위한 검색[6], 구조 정보를 통한 검색기법[8]등이 있다. 이러한 기법들은 xml의 구조적 검색을 지원하고 있으며, 검색의 범위가 확장되고 있다.

2-3. 이름공간

XML 이름공간은 특정한 XML 문서의 종류에 관계된 요소와 속성의 이름과 같은 단순한 이름들의 집합이다. XML어플리케이션을 구성할 때 하나의 간단한 XML문서에 포함된 요소와 속성이 다양한 소프트웨어 모듈에서 사용될 수 있고 정의되어질 수 있도록 하기 위하여 XML 이름공간을 사용하였다. 이와 같은 모듈성은 이해하기가 쉽고 유용한 소프트웨어를 만들어 낼 수 있다. 또한 이러한 마크업 언어는 새로 만드는 것보다 재 사용하는 것이 더 효과적이기 때문에 제안되었다.

따라서 이름공간을 사용함으로써 사용자의 다양한 환경과 사용의 목적에 따라 여러 가지로 사용이 될 수 있으므로 확장성을 보장할 수 있다.

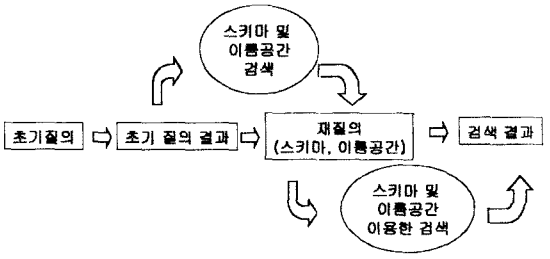
이름 공간을 선언하는 방법에는 두 가지가 있다. 첫째로 디폴트(default) 이름 공간을 선언하는 방법이고 둘째는 접두사(prefixed)이름 공간을 선언하는 방법이다.

3.시스템의 설계

본 논문에서 제시하는 검색 시스템은 구조적 문서 검색시 재질의를 통한 필요한 스키마의 정보를 제공받고, 스카마의 충돌문제와 문서의 사용 목적이나 특성에 따른 자유로운 스키마 확장을 보장하기 위한 이름공간을 이용한 검색을 하였다. 또한 이름공간의 사용으로 인한 정확한 질의의 결과를 얻는 것이 시스템의 설계의 목적이 있다.

3-1.시스템의 질의 과정

질의의 순서는 다름과 같다.



(그림1)질의 과정

3-1-1.초기 질의

초기 질의는 사용자가 스키마와 이름 공간 등을 잘 알지 못하는 상태에서 질의를 하는 것을 가정한다. 일반적인 검색 엔진과 같은 구조로 만들어졌다. 스키마와 이름 공간에 대한 정보에 제한을 받지 않는 keyword 검색이 된다.

예) 홍길동의 저서는 무엇이 있는가?

질의 : 홍길동

3-1-2. 검색의 결과

검색의 결과는 일반적인 검색 기법과는 다르게 엘리먼트와 이름 공간을 표기해 준다. 또한 보다 상세한 스키마 정보가 필요하다면 필요한 문서에 대한 정확한 스키마 구조의 검색이 가능하다. 또한 이름공간에 대한 페이지로 정렬을 하였다. 이름공간이 같다는 것은 스키마의 구조가 같다는 것을 의미하기 때문이다. xml 문서가 웹 브라우저에서 스타일 시트를 적용하지 않은 페이지를 보면 엘리먼트의 구조와 내용이 함께 나타나는 것과 같은 구조를 보여준다. 사용자는 초기 질의 검색결과를 통하여 자신이 원하는 정확한 정보가 어떤 이름공간에 어떤 엘리먼트에 정의되어 있는지를 알게 된다. 그리고 다른 이름공간에 속한 정보와 비교를 할 수 있다.

예)namespace1

<name>홍길동</name> url

namespace2

<n>홍길동</n> url

3-1-3. 재질의

재질의시 검색의 결과를 통하여 스키마와 이름공간을 이용한 정확한 정보 검색 명령을 줄 수 있다. 찾고자하는 정보가 다른 이름공간에 있다면, 서로 다른 구조적 문서의 엘리먼트와 이름공간을 이용하여 검색을 한다.

예)질의: 이름공간, element의 이름과 내용, 찾고자하는 값

3-1-4. 검색의 결과

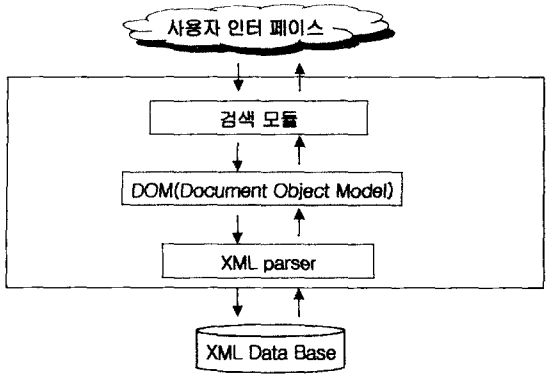
검색의 결과는 임시 저장공간에 저장되어 XSLT를 통하여 추출된 형태의 정보로 표현한다. 검색의 결과를 통합하여 중복된 결과는 삭제를 하고 정렬하여 결과 값을 보여 준다.

예)홍길동의 저서 namespace

달마야 놀자 <url>

엽기적인 그녀<url>

3-2.시스템의 설계



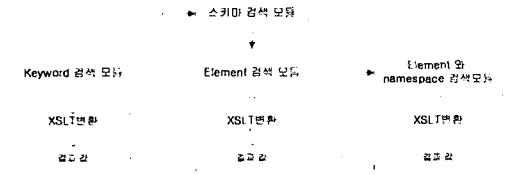
(그림2) 시스템의 구성

시스템은 XML API인 DOM (Document Object Model)을 사용하였다. DOM은 문서의 접근과 조작이 가능하며, 추상화된 계층구조를 갖고 있는 특성 때문이다. 그리고 사용된 언어는 Java와 Javascript이다.

검색 시스템의 모듈은 다음과 같이 3가지로 나누어 구성되어 있다. 사용자가 스키마 및 이름공간에 대한 정보를 알지 못하는 상태에서 질의를 하는 keyword 검색 모듈과 keyword 검색

을 바탕으로한 element 검색, element 검색의 결과를 참조 하여 스키마에 대한 검색, 그리고 이 자료를 통합하여 element와 이름공간을 이용한 검색 모듈로 구성이 된다.

만약 element와 이름공간 등의 정보를 알고 있다면 3단계의 검색이 가능하고, 사용자의 수준별 요구에 따른 검색이 가능하도록 설계하였다.



(그림3)검색 모듈

검색 시스템의 구현을 위하여 XML 스타일 언어인 XSLT(XML Style-sheet Transform)를 사용하였다.

3-2-1.초기 질의 시스템의 설계(keyword 검색)

초기 질의 시스템은 일반적인 검색 시스템과 같다. 하지만 결과를 표기하는 방법이 다르다. 일반적인 검색 시스템은 결과만을 표기하는데 비하여 초기 검색 시스템은 element와 이름공간까지 표시를 해준다.

검색의 결과는 XSLT로 변환된 문서이며 결과의 값은 찾고자 하는 정보가 포함된 element와 이름공간에 대한 정보가 있다.

3-2-2.재질의 시스템의 설계

재질의 시스템은 두 가지로 나눌 수 있다. element 검색만을 이용한 검색과 element와 이름공간을 모두 포함한 검색시스템이다. 이러한 재질의 시스템은 정확한 결과를 검색하여 통합하는 것이므로 몇가지 추가적인 과정이 필요하다.

검색시 element를 검색한 다음 각 element에 대한 이름 공간 검색 모듈이 있고, 검색된 결과를 바탕으로 임시 공간에 저장한 다음 중복을 제거하고 빈도수에 따른 정렬을 하고 출력을 한다.

4.성능분석

4-1. 검색시 충돌의 문제

같은 이름공간을 사용하고 있는 것만 검색의 대상이 되기 때문에 같은 이름을 가지고 다른 의미로 쓰이는 문제의 해결과 같은 의미가 다른 엘리먼트로 쓰이는 것을 비교하여 함께 검색을 할 수가 있다.

4-2. 검색시 정확성

이름공간과 엘리먼트를 사용하기 때문에 구조적인 검색과 함께 정확한 검색이 가능하다.

4-3. 검색의 속도

검색시 재처리하는 과정과 메모리에 모두 로드하는 과정을 통하여 기존의 검색 기법보다는 느리다.

5.결론 및 향후 과제

본 논문에서는 사용자의 편의를 위하여 재질의를 추가 구성하였고, XML 데이터의 자유로운 확장과 정확성을 보장하는 검색 시스템을 설계하였다. 또한 향후 과제로는 여러 개의 이름공간을 통합하여 검색을 할 수 있는 시스템과 데이터의 통합과 확장의 자유를 보장하는 또 다른 방법을 찾아보려고 한다.

참고문헌

(1)T.Bray et "Extensible Markup Language(XML) 1.0 (Second edition)," http://www.w3c.org/TR/2000/REC-xml, 2000.

(2)T.Bray et "Namespace in XML," http://www.w3c.org/TR/1999/REC-xml-names, 1999.

(3)J.clark "XSL Transformations(XSLT) Version 1.0," http://www.w3.org/TR/xslt, 1999.

(4)이승원 권석훈, "XML Schema를 이용한 스키마 통합시 충돌 문제의 분류," 한국 정보과학회, 가을 학술발표논문집, pp.31-33, 2001

(5)A.Deutsch et "A Query Language for XML," http://www.w3c./TR/NOTE-xml-ql, 1998

(6)양승원, 노희영 "시소러스를 이용한 XML 태그 검색 시스템," 한국 정보과학회, 가을 학술발표논문집, 2000

(7)K.Williams etc "XML Databases" 정보문화사, 2001

(8)이성재 "XML 문서를 위한 구조 및 내용 기반 문서 검색 시스템의 설계 및 구현," 한국 정보과학회, 가을 논문 학술논문발표집, pp.93-95, 1999.

(9)문찬호, 강현철 "링크 정보를 활용한 XML 문서의 검색," 한국정보처리학회, 춘계 학술발표 논문집, 2000.

(10)조윤기, 조정길, " XML 문서에 포함된 구조 정보의 표현과 검색," 한국 정보처리학회 논문지, 제8-D권 제 4호, 2001.

(11)J.Stenback etc "Document Object Model Level 2 HTML Specification version1.0," http://www.w3.org/TR/2001/WD-DOM-Level-2-HTML-20011210, 2001

(12)The MITRE Cor, and member of the xml-dev list group XML, "Schemas: Best Practice Homepage" http://www.xfront. com/BestPracticesHomepage.html, 2001

(13)송충범, 유재수, "버저닝을 지원하는 XML 저장관리 시스템 설계 및 구현," 한국 정보과학회, 가을 학술발표논문집, pp.220-222, 2001.

(14)D.florescu and D.kossmann, "Storing and Querying XML data Using an DBMS," Bulletin of Technical Committee on data Engineering, Vol.22, No3, pp27-34, 1999.