구조적 정보 검색을 위한 XQL 질의 처리 시스템 설계

김상영*, 김철원*, 김광현**, 박종훈***, 정현철****
*호남대학교, **광주대학교, ***중부대학교, ****광주보건대

Design of XQL Query Processing System for Structural information retrieval

*Sang-Young Kim,*Chul-Won Kim,**Gwang-Hyun Kim,***Jong-Hun Park,****Hyun-Cheol Jeong
*Honam Univiersity,**Gwangju University,***Joongbu University,****Gwangju health College
E-mail : hrsksy@nate.com

요 약

XML은 단순히 웹 브라우저에 표시하기 위한 것을 넘어서 여러 다양한 시스템간, 어플리케이션간의 데이터 교환을 위한 인터페이스 포맷 등 다양한 분야에서 활용되고 있다. 이에 따라 정보의 생성, 재사용, 처리 및 지속성, 이식성 등 XML 문서를 효과적으로 관리하고 검색할 수 있는 시스템에 관한 많은 연구들이 진행되어 지고 있다. 본 논문에서는 XQL과 문서 구조 처리기와 질의 언어 처리기에 대해 설명하고 XML 문서의 내용을 트리구조로 만들어 구조정보를 파싱하는 동안 XQL을 이용해 질의에 맞는 XML 문서 트리 구조정보를 찾는 방법을 제시한다. 이를 통해 웹 상에 분산된 XML 문서를 병합하여 파싱한 후 문서의 구조 정보를 트리 구조로 구성하고 질의 언어로 제안되어진 XQL을 이용한 효율적인 XML 문서 검색 시스템의 설계 및 구현에 대하여 기술하였다.

ABSTRACT

XML is used in various fields such as interface format for data swapping between application between several various system passing over thing to mark to web browser simply. Accordingly, a lot of studies about system that can manage effectively and search XML document with formation of information, reusability, disposal and durability, portability are proceeding. In this paper, explain about XQL and document structure processor and language processor of quality and make contents of XML document by tree structure, structure information presents method that find XML document tree structure information that is correct on question using XQL while do parsing. Through this, described for design and embodiment of efficient XML document search system that use XQL that compose structure information of document in tree structure and is proposed in language of quality after do parsing absorbing XML document that is scattered on web.

키워드
XPath, XML-QL, XQL, XQuery

Ⅰ. 서 론

XML은 구조화된 문서를 표현하기 어려운 HTML, 그리고 태그의 축약이나 복잡성 등의 문제를 안고 있는 SGML의 단점들을 극복하면서 새롭게 대두된 웹 문서의 표준이다. 데이터로서의 XML은 단순히 웹 브라우저에 표시하기 위한 것을 넘어서 e-Business를 비롯한 MathML, WML 등 여러 분야에서 활용되고 있을 뿐 아니라, 여러 다양한 시스템간, 어플리케이션간의 데이터 교환을 위한 인터페이스 포맷으로도 유용하게 쓰여 질 수 있기 때문에 전자상거래와 국가 행정 업무간 문서 전송과 저장, 자료 검색 등 인터넷과 DB를 사용하는 범위를 뛰어 넘는 다양한 분야로 확장되고 있다. 이에 따라 정보의 생성, 재사용, 처리 및 지속성, 이식성 등 XML 문서를 효과적으로 관리하고 검색할 수 있는 시스템에 관한 많은 연구들이 진행되어 지고 있다.

이와 같이 XML에 대한 관심이 고조되고 실제

많은 분야에서 활용되고 있기 때문에 향후 XML 응용분야의 진행 발전 단계에서 정보의 생성, 재사용성, 처리 및 지속성, 이식성 등과 같은 XML 사용 분야가 점차 증대되고 있다. 또한 DB정보의 양이 급증하고 질의에 대한 사용자 요구가 다양해지면서 보다 정밀하고 신속한 검색을 할 수 있는 정보 검색 시스템 개발이 요구되어 지고 있다.

XML 문서는 기존의 웹 문서와는 달리 내용 정보와 구조 정보를 포함한다. 따라서 기존의 웹 문서에서 제공 되어진 키워드를 통한 내용 정보 검색뿐만 아니라 문서의 논리적인 구조 정보에 대한 검색이 필요하다.

본 논문에서는 내용기반, 구조기반, 속성기반 검색을 지원하는 XQL과 XML 문서를 분석하는 문서 구조 처리기와 사용자 입력 질의를 실행하기 위한 질의 언어 처리기에 대해 설명하고 XML 문서의 내용을 트리구조로 만들고 구조정보를 파싱하는 동안 XQL을 이용하여 질의에 맞는 XML 문서 트리 구조정보를 찾는 방법을 제시한다.

이를 통해 웹 상에 분산된 XML 문서를 병합하여 파싱한 후 문서의 구조 정보를 트리 구조로 구성하고 질의 언어로 제안되어진 XQL을 이용하여 JAVA 기반의 효율적인 XML 문서 검색 시스템의 설계 및 구현에 대하여 기술하였다.

논문의 구성은 2장에서 관련 연구를 기술하고, 3장에서는 검색 시스템 설계 및 구현을 보이며 4장에는 결론 및 향후 연구 과제를 설명한다.

II. 관련연구

2.1 XML 문서 저장 및 검색 시스템

XML 문서 저장 및 검색 시스템은 XML 문서의 구조정보를 유지하면서 효율적으로 저장 및 검색, 관리할 수 있는 시스템을 말한다.

XML 문서는 데이터 중심의 XML과 문서중심의 XML로 구분되어 진다. 데이터 중심 XML문서는 정형적인 구조로써 내용과 구조가 혼합되어 있는 양이 적어 어플리케이션과 데이터 저장소간의 데이터 교환을 위해 사용 되어진다. CALS/EC 등 대규모의 표준화된 데이터, 비행 스케줄, 메시지 등이 데이터 중심 XML 문서이다. 반면 문서중심의 XML은 비정형적인 구조로써 엘리먼트와 내용이 혼합된 형태를 가지므로 전자도서관(Digital Library)이나 전자북(eBook), 계약서, 전자 매뉴얼(ITEM), 광고 등과 같은 문서를 생성하는데 사용되어진다.[2]

기존 XML 문서 저장 및 검색 시스템에 대한 연구로는 XML Database System에 해당하는 eXcelon이나, Oracle 8i, Tamino 등이 있으며 XML Content Management System으로는 Bladerunner, POET CMS등이 있다. eXcelon은 단순성과 유연성과 같은 XML 장점을 살릴 수 있는 응용 프로그램을 개발하도록 하는 XML 데이터 서버이다. W3C의 XML 표준틀을 적용하여 문서단위의 입출력뿐만 아니라 XML 노드 레벨의 저장, 검색이 가능하며 인덱스를 생성하여 빠르고 정확한 검색을 할 수 있다. Tamino는 XML 데이터를 별도의 변환과정 없이 표준 XML 포맷으로 저장하는 스토리지, RDB, 계층형 데이터베이스 등 기존의 주요 데이터 소스에 접근 위한 인터페이스, SQL 데이터를 위한 스토리지, 웹상에 분산되어 있는 Tamino DB 서버들에 대한 중앙 관리 도구를 제공하는 XML 전용 데이터베이스이다.

2.2 XML 질의 언어(XML Query Language)

XML 질의 언어의 종류로는 XML문서의 패턴 검색, XSL 패턴 문법에서 발전 되어진 XQL(1998), Web 환경에서 다량의 XML 데이터의 사용성을 높이는 XML-QL(1998), 그리고 XQL과 XML-QL의 특징을 통합한 XQuery(2001)등이 있다.

XQL은 XML 문서로부터 문서 또는 특정 엘리먼트/텍스트를 접근하거나 검색하기 위한 질의 언어이다. XSL을 확장한 형태로 질의와 패턴을 사용하는데 단순한 디렉토리 표기법이나 Boolean 표현 등(=, !=, eq, ne,..)을 사용 하므로 특정 노드 또는 엘리먼트를 명시하는 정확한 표기가 가능하다.[3]

XML-QL은 다량의 XML 문서에서 데이터를 추출, 한 문서에서 다른 문서로 변환, 여러 장소의 XML 데이터 통합에 대해 정의된 언어이다.[4] XML-QL은 패턴을 사용하여 데이터를 매칭하며 결과로서 새로운 XML 데이터를 생성하게 된다.

Quilt[5]에서 유래된 XQuery[6]는 다른 언어들이 가지고 있는 몇가지 유용한 기능들을 추가한 것으로서 XML 구조를 이용하여 전체가 구조화되거나 부분적으로 구조된 XML 문서에 대한 복잡한 질의를 수행할 수 있다.

본 논문에서는 XML 검색 질의어로 XQL을 사용하여 검색할 수 있도록 한다.

III. XML 문서 검색 시스템 설계 및 구현

본 논문의 문서 검색 시스템은 자바 언어로 구현하였고, JDK1.4, JSP, SUN사의 DOM과 SAX를 지원하는 파서와 XML 질의어로는 XQL을 기반으로 설계하였다. <그림1>은 본 논문에서 제안한 XML 문서 검색 시스템의 전체적인 구성도 이며 시스템의 구조는 크게 문서 처리부분과 XQL 질의 처리부분으로 나누어 볼 수 있다.

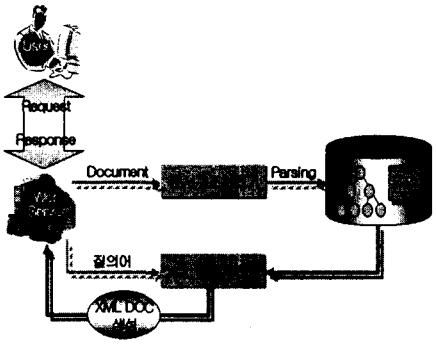

그림 1. XML 문서 검색 시스템 구성도

이 시스템에서는 XML 문서의 엘리먼트, 속성과 내용에 대하여 검색 할 수 있도록 시스템을 구성하였고, 그 질의 결과로는 XML 문서를 출력할 수 있도록 하였다.

3.1 XML 문서 검색 시스템 구성

XML 편집기에 의해 작성된 문서의 내용은 파싱한 후 JDBC/ODBC 인터페이스에 의해 데이터베이스에 저장을 한다. 질의어 처리 과정은 Well-formed XML 문서를 입력 받아서 문서 구조 처리기에 의해 파싱한 후, start/end document, start/end element, 텍스트들을 분류하여, XML 문서를 인덱스와 문서구조 정보를 구성하고 그 정보를 XML 질의어 처리기의 문서 검색을 위해 저장한다. 질의 결과에 대한 구조 정보는 인덱스 정보와 함께 스택에 유지되어지며, 결과에 대한 질의를 계속하여 처리할 수 있도록 인덱스 정보를 계속 유지한다.

3.2 문서 구조 처리기

문서 구조 처리기는 XML 문서를 파싱하면서 인덱스 구조정보를 생성시킨다. Well-formed XML 문서를 입력받아 스택과 구조정보 및 텍스트 내용 정보 등에 관하여 초기화 한 후, start/end document, start/end element, 텍스트(내용)들을 분류하여, XML 문서를 인덱스된 데이터 구조로 변환시킨다. 트리 진행순서는 입력된 XML 문서의 순서에 따라 진행하고, 스택에 의해 저장 되어진다. 이것은 XML 문서의 순서와 일치한다.

<그림2>는 각 노드에 따라 구조화 정보를 결정하면, 인덱스 및 구조정보를 가진 데이터 구조를 생성한다. 각 노드는 엘리먼트를 나타내며, 노드의 숫자는 레벨에 따른 부모노드와 자식노드, 형제노드와의 순서를 갖고 있다. 속성 검색을 위해 속성은 내부 인덱스 정보를 구성하고 있으며, 텍스트 검색을 위한 단어 단위의 인덱스 정보를 갖고 있다.

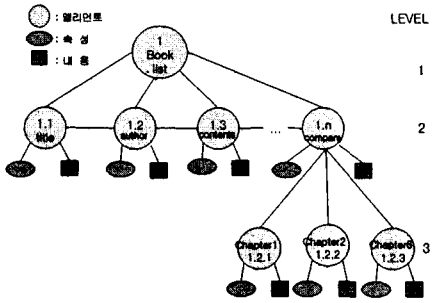

그림 2. 각 노드의 인덱스 및 구조 정보

3.3 XQL 질의어 처리기

XQL 질의어 처리기 에서는 사용자로부터 질의어를 입력받고, 질의어에 대한 구문분석을 처리한 후 엘리먼트 검색, 속성검색, 컨텐트 검색에 대하여 분류하고 기존의 인덱스정보와 구조정보를 구성하였던 인덱스 문서구조 파일을 참조하여 검색한다. 속성검색의 경우 속성 이름과 속성 값에 대하여 인덱스정보를 검색하며, 컨텐트 검색의 경우 단어 단위로 인덱스된 정보를 검색한다.

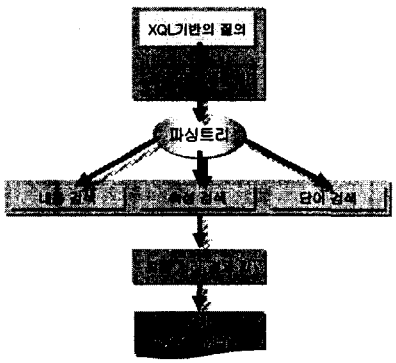

그림 3. XQL 질의 처리기

질의어 처리기 순서로는 XQL 질의어가 입력되면 XQL 질의어 구문 분석기로 파싱해서 파싱 트리를 생성하게 되고, 그 후 파싱된 트리를 기존의 인덱스 정보 및 구조 정보로부터 엘리먼트 검색, 속성 검색, 컨텐트 검색을 한 후 최종 처리 결과 값을 XML 문서로 출력하게 된다.

3.4 질의어의 예와 출력 결과

3.4.1 입력된 XML 문서

<그림 4>의 Booklist를 최상위 엘리먼트로 가지는 XML 문서를 이용하여 XQL 질의어에 대한 검색을 한다. 간단한 예로 Booklist 엘리먼트 한가지만 제시하였다

```
<?xml version="1.0" encoding="EUC-KR"?>
<Booklist>
<title num="34" category="computer">
XML Data Management : Native XML and XML-Enabled Database
Systems</title>
<author>Akmal B. Chaudhri , Roberto Zicari , Awais Rashid </author>
<company>Addison-Wesley</company>
<publish>20030312</publish>
<image>computer34.gif</image>
<price>40850</price>
<contents>
<capter1>1. Information Modeling with XML</capter1>
<capter2>2. Tamino – Software AG's Native XML Server</capter2>
<capter3>3. eXist Native XML Database</capter3>
<capter4>4. Embedded XML Databases</capter4>
....
</contents>
<title num="34" category="computer">
....
</Booklist>
```

그림 4. 입력된 XML 문서

3.4.2 질의어 예와 출력 결과 문서

다음 그림은 "*booklist1.xml*"라는 XML 문서를 입력으로 엘리먼트 검색, 속성 검색, 내용검색으로 분류하여 질의한 출력 결과 화면이다.

1) 구조/내용 기반 검색

Query : //Booklist/title[. &= "XML"]

설 명 : Booklist 엘리먼트를 먼저 찾고 그 자식 노드인 title 엘리먼트의 내용 중에서 'XML'이 포함된 노드만을 출력한다.

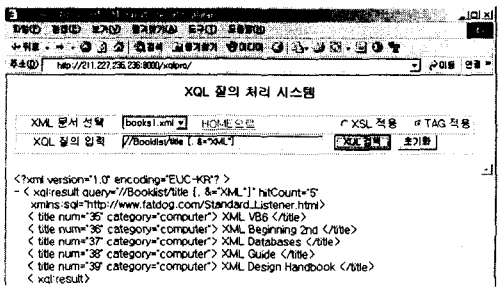

그림 5. 내용 검색 출력결과 화면

2) 속성 기반 검색

Query : //Booklist/title[@num = "2"]

설 명 : Booklist 엘리먼트에서 그 자식 노드에 있는 Booklist 엘리먼트 중에서 속성 이름이 id이며 값이 1인 title 엘리먼트들을 출력한다

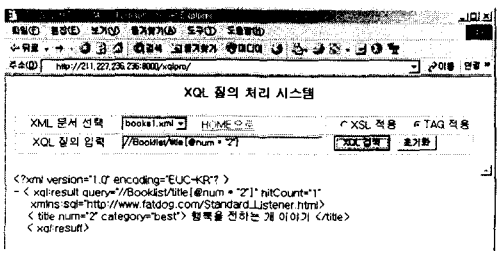

그림 6. 속성 검색 출력 결과 화면

3) 엘리먼트 검색

Query : //Booklist/title

설 명 : Booklist 엘리먼트의 자식 노드인 title 엘리먼트들을 모두 출력한다.

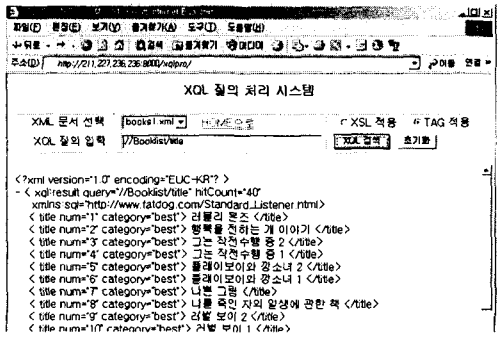

그림 7. 엘리먼트 검색 출력 결과 화면

IV. 결론 및 향후과제

XML은 웹 상에서 상업 데이터 교환을 위한 표준으로 기업간(B2B), 또는 기업과 정부간(B2G), 기업과 소비자간(B2C)과 같은 전자상거래에서 정보교환을 위한 문서 포맷으로 사용되어지고 있다. 이처럼 문서의 처리 및 이 기종 시스템간의 정보 교환은 그 중요성이 계속 증가되고 있으며, 이에 대한 XML 문서에 대한 저장과 검색이 점차적으로 중요해 지고 있다. 따라서, 현재 XML 분야에 자체 기술 축적을 위해 본 논문에서는 XML 문서에 대한 구조와 표준 질의로 검색할 수 있는 시스템을 설계하였다. Well-formed XML 문서를 입력으로 하고 XML 문서구조 처리기에 의해 XML문서를 구조 분석하고, XQL 기반의 질의어 처리기에 의해 문서의 엘리먼트와 속성, 내용을 검색할 수 있도록 구성하였다.

본 논문에서 구현한 XML 문서 검색 시스템은 B2B, ECommers, e-Book, XML/EDI등 과 같은 여러 응용분야에 광범위하게 적용될 수 있는 기술로 예상된다.

향후 연구로는 질의 입력 방식을 개선하여 XQL Query 문법을 잘 알지 못하는 최종 사용자도 쉽고 간편하게 사용할 수 있도록 interface를 개선해야 할 것이며, 온라인상에서 대규모의 XML 데이터를 효율적으로 이용하면서 분산된 XML 문서를 저장 및 관리할 수 있는 환경으로 확장시킬 필요가 있다. 뿐만 아니라 질의 결과를 사용자가 원하는 형태의 문서 포맷이나 다른 형태의 컨텐츠로 제공해야 할 필요성이 있다.

참고 문헌

[1] Extensible Markup Language(XML)1.0, W3C Recmmendation, http://www.w3.org/TR/1998/REC-xml-19980210, 1998.

[2] Ronald Bourret, "XML and Databases", http://www.rpbourret.com/xml/XMLAndDatabases.htm, 2003.

[3] J. Robie, J. Lapp, D. Schach. XML Query Language (XQL)http://www.w3.org/TandS/QL/QL98/pp/xql.html

[4] A. Deutsch et al., "XML-QL : A Query Language for XM," http://www.w3.org/TR/NOTE-xml-ql/,1998

[5] D.Florescu et al. "Quilt : An XML Query Language for Heterogeneous Data Sources", http://www.almaden.ibm.com/cs/people/chamberlin/qult_incs.pdf, 2000

[6] "XQuery 1.0: An XML Query Language", http://www.w3.org/TR/2002/WD-xquery-20021115/, 2002