제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

XML 데이타의 효율적인 검색을 위한 색인 모델에 관한 연구

권국봉, 홍동권
계명대학교 정보통신대학
e-mail:{gbkwon,dkhong}@kmu.ac.kr

A Study on Indexing Technique for Efficient Search of XML Data

Guk-Bong Kwon, Dong-Kweon Hong
College of Information and Communication, Keimyung Univ.

요 약

본 연구에서는 XML 데이타의 효율적인 검색을 위한 색인 모델을 제안한다. 제안한 색인 모델은 문서 계층상의 모든 레벨에서 내용 기반 질의, 구조 기반 질의와 같은 다양한 질의를 지원하기 위한 구조정보와 이를 이용한 색인 구조로 구성된다. 그리고 구조 검색을 지원하기 위해 새로운 구조정보 표현 방법을 제안한다. 또한 제안된 색인 모델에 지속성을 부여하기 위해 색인 모델을 디스크에 저장하는 방법을 제안하고 간단한 질의의 처리 과정을 설명한다.

1. 서론

최근 정보의 양이 급증하면서 정보를 보다 효율적인 방법으로 표현하고, 사용하고자 하는 연구가 활발히 진행되고 있다. 한편 현재 가장 많이 사용하고 있는 기술로는 월드 와이드 웹(World Wide Web)을 말할 수 있는데 웹의 가장 기본 기술이 되는 HTML(Hyper Text Markup Language)는 정보의 표현에 중점을 두고 있어 정보의 저장 및 검색하는 기능은 부족하다는 단점이 있다. 이에 웹 발전에 주도적인 역할을 하고 있는 W3C(World Wide Web Consortium)에서는 차세대 웹 문서의 표준으로 XML(eXtensible Markup Language)이라는 전자문서 메타언어를 1996년에 제안하였으며 현재까지 그 기능이 계속 확장되고 있는 상태이다.[1]

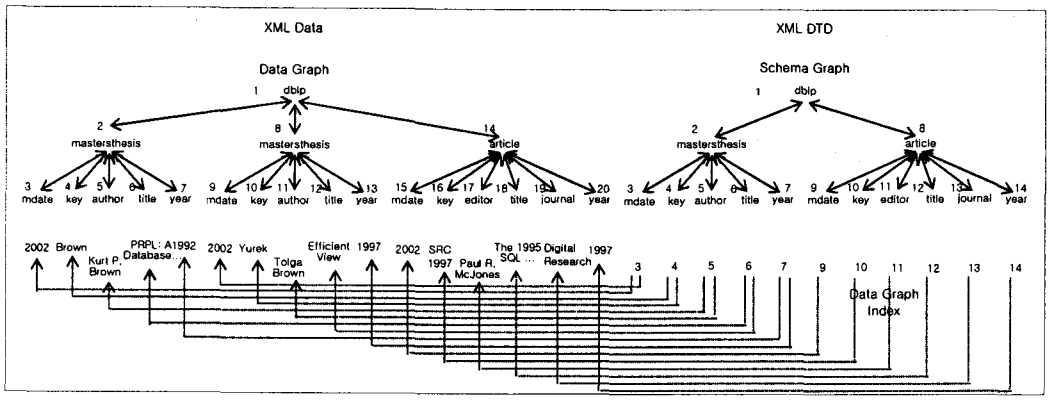
XML은 이질적인 시스템에서 작성된 문서의 상호 교환과 다양한 형식의 문서들을 일관성 있게 구조화하기 위해 고안된 SGML을 간략화 시키고 HTML보다 사용자의 다양한 요구를 충분히 수용할 수 있어 웹 문서뿐만 아니라 전자도서관, e-business등 다양한 분야에서 XML을 활용하고자 폭넓은 연구를 하고 있으며 구체화되는 XML 응용이 많아지고 있다.

XML의 이러한 목적은 데이타베이스 관리 시스템의 목적과도 비슷하게 볼 수 있다. 그러나 현재 사용되고 있는 데이타베이스 관리 시스템은 현실 세계의 복잡한 형태의 데이타들을 쉽게 표현하지 못할 뿐만 아니라 저장된 데이터들도 데이타베이스 관리 시스템의 차이로 이질적인 시스템 사이의 데이타 교환을 잘 지원하지 못한다.

한편 XML의 이러한 구조적 유동성은 모든 형태의 데이타가 XML로 기술될 수 있는 기반을 제공한다.

앞으로 많은 데이타가 XML로 표현 될 것이고 이런 데이타를 효율적으로 관리하기 위해서는 데이타

※ 본 연구는 한국과학재단 목적기초연구 (R01-2003-000-10001-0)지원으로 수행되었음.



[그림 1] 색인 모델의 전체 구조

베이스 관리 시스템의 사용이 필수적이다. XML 데이타를 관리하기 위한 방법은 기존에 많이 연구되어 왔는데 크게 두 가지로 나누어 볼 수 있다. 한가지는 XML 데이타를 위해 새로운 구조의 데이타베이스 관리 시스템을 만드는 것이고, 다른 한가지는 기존의 데이타베이스 관리 시스템에 XML 데이타를 저장하는 것이다.

이에 본 논문에서는 새로운 구조의 XML 데이타베이스 관리 시스템을 위해 DTD나 스키마에 나타난 XML 데이타의 구조 정보를 사용해서 XML 데이타를 효율적으로 관리하고 검색 할 수 있는 데이타 모델을 제안하고 이 데이타 모델을 이용해 효율적인 검색을 위한 색인 구조를 구성하는 색인 모델을 제안한다. 나아가 제안된 색인모델에 지속성을 부여 위해 디스크에 저장하는 방법에 대해서 연구한다.

본 논문의 구성은 다음과 같다. 2절에서는 데이타베이스 관리 시스템과 XML의 관계에 대해서 알아보고 3절에서 효율적인 검색을 위한 구조정보 표현을 제안하고 전체적인 색인 모델을 제안한다. 4절에서는 제안된 색인 모델의 지속성 부여 방법에 대해서 제안하고 5절에서는 제안된 색인 모델을 이용해 간단한 XML 데이타의 색인을 구성하고 간단한 질의 처리의 예를 보인다. 마지막으로 6절에서는 결론과 향후 연구 방향을 제시한다.

2. 데이타베이스 관리 시스템과 XML

데이타베이스 관리 시스템은 크게 두 가지 목적에 사용되는데 하나는 데이타의 저장과 검색이고 다른 하나는 이질적인 시스템 사이의 데이타의 교환 및 중계를 위해 사용한다. 그러나 현재 사용되고 있는 데이타베이스 관리 시스템은 현실 세계의 복잡한 형태의 데이타들을 잘 표현하지 못할 뿐만 아니라 저장된 데이타들도 데이타베이스 관리 시스템의 차이로 이질적인 시스템 사이의 데이타 교환 및 중계를 잘 지원하지 못한다.

한편 XML은 DTD나 스키마를 통하여 데이타 자체에 데이타의 구조를 기술하고 있다. 이와 같이 데이타의 구조는 사용자가 원하는 대로 정의할 수 있으며, 이러한 구조적 유동성은 모든 형태의 데이타가 XML로 기술될 수 있도록 해주고 이질적인 시스템에 사이의 데이타 교환을 잘 지원한다. 이것은 현실 세계에서 운용되는 모든 데이타가 동일한 형태로 통합, 저장, 처리 될 수 있는 기반을 제공 해 준다.

3. 색인 모델

[그림 1]과 같이 본 논문에서 제안하는 색인모델은 크게 XML 데이타를 표현하는 Data Graph, XML 스키마를 표현하는 Schema Graph, XML Document Graph를 색인 하는 Index 등 3가지로 구성되어 있다.

```
<!ELEMENT dblp (article|mastersthesis)*>
<!ENTITY % field "author|editor|title|year
                  |journal">
<!ELEMENT mastersthesis (%field;)*>
<!ATTLIST mastersthesis key CDATA #REQUIRED
                        mdate CDATA #IMPLIED>
<!ELEMENT article (%field;)*>
<!ATTLIST article key CDATA #REQUIRED
                  reviewid CDATA #IMPLIED
                  rating CDATA #IMPLIED
                  mdate CDATA #IMPLIED>
<!ELEMENT author  (#PCDATA)>
<!ELEMENT editor  (#PCDATA)>
<!ELEMENT title   (%titlecontents;)*>
<!ELEMENT year    (#PCDATA)>
<!ELEMENT journal (#PCDATA)>
<!ENTITY % titlecontents "#PCDATA|sub|sup
                          |i|tt|ref">
```

[표 1] XML DTD

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
  <mastersthesis mdate="2002" key="Brown">
    <author>Kurt P. Brown</author>
    <title>PRPL:A Database Language</title>
    <year>1992</year>
  </mastersthesis>
  <mastersthesis mdate="2003" key="Yurek">
    <author>Tolga Yurek</author>
    <title>Efficient View</title>
    <year>1997</year>
  </mastersthesis>
  <article mdate="2002" key="SRC1997">
    <editor>Paul R. McJones</editor>
    <title>The 1995 SQL Reunion</title>
    <journal>Digital Research</journal>
    <year>1997</year>
  </article>
</dblp>
```

[표 2] XML 데이타

[그림 1]은 [표 2]의 XML 데이타를 제안하는 색인 모델로 표현 한 것으로 크게 세 부분으로 이루어 져 있다.

첫째, [그림 1]에서 왼쪽 그래프는 XML 데이타를 Data Graph로 변환한 것을 보여 주는데 XML 데이타의 각 엘리먼트는 Data Graph의 하나의 노드에 해당한다. 트리는 XML 데이타의 부모 및 자식의 구조 정보를 그대로 유지하며 각 노드는 트리를 깊이 우선 탐색 기법으로 탐색하는 순서로 고유한 노드 아이디를 가지게 되고 이 노드 아이디는 나중에 Schema Graph 단말 노드에 있는 Data Graph의 인덱스를 생성할 때 기록되는 노드의 고유한 정보이다. 또한 각 노드에는 부모의 노드 아이디 , 자식의 노드 아이디들 등 각 노드들의 구조에 대한 정보를 노드 아이디로 구분하여 가지게 된다.

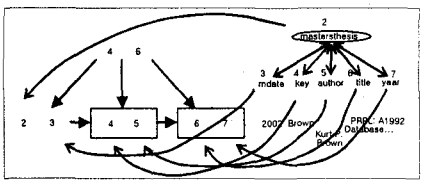
둘째, XML DTD를 표현하는 Schema Graph는 [그림 1]에서 오른쪽 그래프에 해당하는데 XML DTD의 각 엘리먼트는 Schema Graph의 하나의 노드에 해당한다. Schema Graph는 Data Graph와 동일한 모양으로 작성되나 Data Graph와 달리 단말 노드에 Data Graph의 인덱스를 가지고 있다.

셋째, Data Graph를 색인 하는 Index는 Schema Graph의 단말노드에 각각 연결되어 Schema Graph의 노드 구조와 일치하는 Data Graph의 노드를 색인 하여 B+트리로 구성한다.

4. 색인 모델의 지속성

앞서 제안된 XML 데이타 색인 모델은 메인 메모리에서 트리 기반의 자료 구조로 구현이 된다.

따라서 제안된 색인 모델이 XML 데이타의 색인으로 사용되기 위해서는 지속성을 가져야 하는데 본 논문에서는 색인 모델의 지속성을 위해서는 디스크에 저장하는 방법을 제시한다.

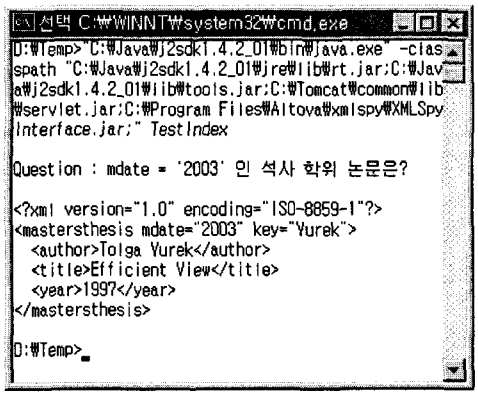


[그림 2] 그래프를 B+트리에 저장

메인 메모리 상의 자료 구조를 디스크에 저장하는 가장 좋은 방법은 트리의 높이를 가장 낮게 유지해서 디스크 접근회수를 낮게 유지 시켜 주는 B트리 계열의 자료 구조를 사용하는 것이다. 앞서 제안된 색인 모델에서 Data Graph Index는 B+ 트리로 구성이 되어 있기 때문에 바로 디스크에 저장 할 수 있고 나머지 Data Graph, Schema Graph를 디스크 저장하기 위해서 [그림 2]에서와 같이 트리의 노드 아이디를 고유한 키로 구분하여 B+트리의 단말 노드에 각 노드를 사상하여 B+트리로 저장한다.

5. 색인 모델의 구현

제안된 XML 데이타의 색인 모델은 Windows 2000 Professional 환경에서 개발되었다. 자바 언어로 개발되었기 때문에 이식성과 플랫폼 독립성을 제공하며, 자바 기술을 위해서 썬 마이크로 시스템의 JDK 1.4.2를 기반으로 하였다. 사용하는 주요 패키지로는 XML 데이타를 파싱하고 조작하기 위해 Apache Software Foundation의 Xerces와 JDOM을 사용하였고 B+트리의 구현을 위해 GiST 패키지를 사용하였다.[3] 예제로 사용한 XML 데이타 파일은 약 14만 라인으로 이루어지고 약 6.5M의 크기를 가지며 이를 색인 모델로 구축하여 파일로 저장했을 때 전체 약 7M의 크기를 보였다.



[그림 3] 예제 프로그램 실행

[그림 3]은 제안한 색인모델을 예제 XML 데이타에 적용하여 색인을 만든 후 간단한 질의를 실행한 결과이다. 질의의 수행 과정은 XML 데이타에 대한 질의가 수행이 되면 먼저 Schema Graph에서 구조 정보를 파악하여 찾아야 하는 정보의 인덱스을 얻는다. 그런 다음 인덱스가 가르키고 있는 Data Graph의 노드를 찾고 그 노드에서부터 시작하여 부모, 자식 노드들에 대한 정보를 얻어 사용자가 원하는 결과를 보여 준다. 예를 들면 예제 XML 데이타에 대해 [mdate = "2003"인 석사학위논문]라는 질의를 수행하면 맨 먼저 Schema Graph에서 mastersthesis 노드에서 mdate 노드를 찾고 노드에 연결되어 있는 Data Graph Index를 얻어서 Data Graph에서 mdate 노드의 아이디를 얻은 후 조건 mdate가 2003년인 노드에 대해서만 노드의 정보를 출력한다.

6. 결론

본 논문에서는 XML 데이타에 대해 효율적인 검색을 위한 색인 모델을 제안하였다. 제안한 색인모델은 크게 세 부분으로 나뉘어 지는데 XML 데이타를 표현하는 Data Graph와 XML 데이타의 구조 정보를 표현하는 Schema Graph, 그리고 Data Graph를 색인 하는 Data Graph Index로 이루어진다.

Data Graph는 XML 데이타의 구조 정보를 잘 표현할 수 있는 트리 형태로 표현되고 이를 깊이 우선 탐색 기법을 사용하여 방문하면서 각 노드마다 고유한 노드의 아이디를 부여하며 이를 사용하여 노드의 구조 정보를 색인 한다. Schema Graph 는 XML 데이타의 DTD를 사용하여 XML 데이타의 구조를 트리 형태로 표현하고 각 단말 노드에 구조 정보와 일치하는 Data Graph의 노드 아이디 색인을 연결하여 Data Graph 탐색이 가능하게 한다. 마지막으로 각 Graph마다 노드 아이디를 키로 사용하는 B+트리에 각 노드를 사상하고 B+트리를 디스크에 저장하여 이들 색인 모델에 대해 지속성을 부여하였다.

이와 같이 제안한 색인 모델을 통해 특정 엘리먼트에 대한 직접적인 접근이 가능하며, 다양한 구조적 질의를 효과적으로 처리 할 수 있다. 따라서 보다 효율적이고 빠른 검색을 지원할 수 있게 되었고 현실 세계에서 운용되는 모든 데이타가 동일한 형태로 통합, 저장, 처리 될 수 있는 기반을 제공하였다.

향후 연구로서 현재 구현되어 있는 색인모델의 지속성 유지 방법에 대해서 연구하고 나아가 XML 데이타의 갱신이 발생하는 동적인 환경에 본 논문에서 제안한 색인 모델을 적용하기 위한 연구가 필요하다. 나아가 제안된 색인 모델을 이용하여 XML 데이타의 질의 최적화 할 수 있는 질의 최적화와 관련된 연구가 필요하다.

참고 문헌

[1] Extensible Markup Language(XML) 1.0 "http://www.w3.org/TR/REC-xml"

[2] W3C. document Object Level(DOM) Level 1 Specification, http://www.w3.org/TR/, Oct 1998

[3] A Generic Indexing Mechanism For Persistent Java, "http://people.cs.uct.ac.za/~evoges/web/"

[4] Jason McHugh Jennifer Widom "Query Optimization For Xml", Proceedings of 25th International Conference on Very Large Data Bases, 315-326, 1999

[5] Leela Krishna Poola Jayant R. Haritsa, "SphinX: Schema-conscious XML Indexing", Database Systems Laboratory Dept. of Computer Science Automation Indian Institute of Science, 2001

[6] 한성근 외 4명, "동적 환경에 적합한 SGML 인덱스 관리자의 설계 및 구현". 한국정보처리논문지, 제6권 제 10호, 1999

[7] 박상원외 3명, "XML과 데이터베이스" , 한국정보과학회지, 제19권 제1호, 2001

[8] 조윤기외 3명, "XML 문서에 포함된 구조 정보의 표현과 검색" , 정보처리학회논문지, 제8권 제4호, 2001