XML 문서의 구조 검색을 위한 저장 시스템 설계 및 구현

정병인* · 김희준 · 이재완
군산대학교 전자정보공학부

A Design and Implementation of an XML Document Storage System for Structural Query

Byoung-in Jeong* · Hee-jun Kim · Jae-wan Lee
*Kunsan University
E-mail : indan@dslab_kunsan.ac.kr

요 약

정보 통신 기술의 발전에 따라 복잡하고 다양한 정보들을 컴퓨터를 이용한 합리적인 방법을 통해 저장, 관리, 검색하여 활용해야 할 필요성이 높아지고 있다.

이러한 환경에서 많이 이용되고 있는 문서 포맷 중의 하나가 XML 이다. XML은 SGML과 HTML의 단점을 해결 및 보완한 것이기 때문에 인터넷을 기반으로 하여 많은 분야에서 활용이 될 전망이다. 따라서 증가하는 XML 문서들을 저장하고 관리하는 기능은 필연적으로 필요하게 된다. 또한 많은 양의 저장된 문서들에 대해서 원하는 문서를 찾을 수 있는 효율적인 검색 기능이 필요로 하게된다.

따라서 제안한 XML 저장관리 시스템은 XML이 지닌 다양한 문서 정의들에 대한 손실이 없이 저장할 수 있는 모델과 데이터베이스에 최소한의 부하만을 주어 구조기반 검색을 수행할 수 있는 검색기 등을 설계 및 구현하였다.

Ⅰ. 서 론

현재까지의 웹과 인터넷의 발전에 가장 큰 공헌을 한 것 중 하나가 바로 HTML(Hypertext Mark-up Language)이다. 그러나 HTML은 쉽게 사용할 수 있다는 장점이 있지만 단순하고 고정된 태그(tag)만을 사용하여 확장에 한계가 있다. 그리고 다른 마크업 언어인 SGML(Standard Generalized Mark-up Language)은 강력한 표현능력을 가지지만 그 문법이 너무 어려워 구현이 어려운 단점이 있다. 이에 최근 W3C(World Wide Web Consortium)는 HTML의 한계를 극복하고 SGML을 단순화하여 보다 사용하기 쉬운 차세대 언어 XML(eXtensible Markup Language)을 제안하였다[1].

XML은 SGML의 복잡성과 HTML의 단순성을 극복하였으며, 상호운용성과 편리한 구현환경을 제공하고, 구조적인 문서구조를 가져 향후 인터넷 및 전자도서관, 문서정보시스템, 전자상거래에서 표준 문서처리 언어로 주목받고 있다.

XML 문서는 기존의 문서와는 달리 하나의 문서에서 내용정보와 함께 문서의 구조정보 즉 저자, 제목, 서론, 본문과 같은 문서의 논리적인 구조 정보를 지니고 있다. 따라서 기존의 문서에서 제공하던 내용정보에 대한 검색 뿐만 아니라

XML 문서의 검색에서는 이러한 논리적인 구조 정보에 대한 검색 기능도 필요하게 된다 본 연구에서는 XML의 특징인 특정 문서종류에 의존되지 않으며 문서구조 정보를 활용하여 문서를 검색하는 문서 검색 시스템을 설계 및 구현하고자 한다.

2. 관련연구

2.1 XML 개요

XML은 SGML에서 중요하게 사용하지 않는 것을 축소 또는 삭제하고 꼭 필요한 기능만 수용하였으며 문서의 논리적인 구조정보의 기술을 간편하게 하였다. 그리고 XML 인스턴스를 만들기 위한 첫 번째 작업은 만들고자 하는 문서의 논리적 구조를 표현할 수 있는 DTD(Document Type Definition)를 정의할 수 있으며, 이 DTD는 HTML과는 달리 고정되어 있지 않으므로 다양한 논리적 구조를 표현할 수 있는 유연성을 갖는다. 즉 어떤 문서의 구조를 기술하는 방법과 그 문서 구조에 맞게 내용을 생성하는 방법을 제시할 수 있다.

XML의 몇 가지 특성을 정리해보면 다음과 같다.

1. 간결성
2. 웹에서 일반화된 마크업을 지원한다.
3. SGML의 규범에 따라 이상적으로 유효화될 수 있는 문서를 만든다
4. 확장성(태그에 대한 제한이 없음)
5. URL접근법과 이상적으로 호환될 수 있는 하이퍼 링킹을 지원한다
6. 문서의 재사용이 용이
7. 플랫폼, 응용에 독립적
8. 웹사이트들의 간단한 시스템 관리 제공
9. 일반적이고 강력한 스타일 메커니즘을 제공

2.2 데이터 검색 모델

논리적인 구조를 가지고 있는 데이터 모델에 대해 구조정보를 이용하여 검색을 수행하는데 활용하기 위한 기법으로 구조화 문서의 색인 기법과 질의 언어 기법(XQL(XML Query Language), XML-QL)에 대한 연구가 진행되고 있다[5].

질의 언어 기술은 문서의 구조적인 특성을 반영한 구조, 내용, 속성 기반 검색을 지원하고 있다. 또한 구조화된 문서의 색인 기법은 파스트리 상에서 노드에 따라 색인어를 중복하거나 중복을 배제하는 기법들을 제시하고 있으며, 각 기법들의 검색 성능을 비교하고 있다[4].

그렇지만 색인어를 데이터베이스에 함께 저장함으로써 저장공간의 효율성을 떨어뜨리며, 구조가 복잡해짐에 따라 테이블 수가 증가하는 단점을 가지고 있다.

3. 시스템 구성 및 DB 저장

3.1 시스템 구성

본 시스템의 개발 목적은 논문 데이터를 공통된 포맷을 사용하여 데이터베이스화하여 문서처리의 비용절감 및 효과적인 데이터 검색을 목적으로 한다.

하나의 문서에 속하는 구성요소들을 불러들여 다시 데이터베이스는 XML 포맷으로 이미지 및 텍스트 그리고 테이블이 복합적으로 혼재하고 있는 문서형태를 원활하고, 효율적으로 관리할 수 있게 설계되었으며 또한 검색을 위한 질의화면이 제공되어 사용자가 쉽게 웹상에서 작업 할 수 있도록 설계되었다.

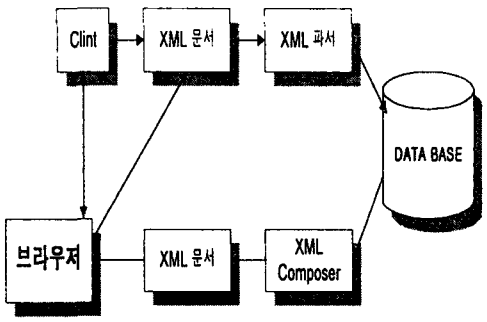

그림 1 XML 문서 처리 과정

그림 1은 XML 문서를 처리하는 과정을 나타낸 것이다. 파서는 입력된 XML 문서를 태그별로 분리하고 분리된 XML 문서의 구성요소들은 각

각의 순서, 위치 및 내용 정보와 함께 DB에 저장된다.

조정자(composer)는 각각의 정보들을 종합하여 XML 문서를 작성한다.

3.2 DTD 설계

XML 문서는 자신의 DTD를 가질 수 있는데, 서로 다른 종류의 DTD의 문서라도 같은 시스템에서 저장되고 관리되어야 한다. 그래서 기존의 저장시스템처럼 DTD 별로 스키마를 새로 생성하지 않는다.

그리고 각 엘리먼트들은 특정한 속성들을 가질 수 있는데, 그러한 속성정보를 저장하기 위하여 속성테이블을 만들어 속성들을 저장한다.

본 시스템에서 논문 데이터 처리를 위한 문서의 정의를 위하여 사용한 DTD는 아래와 같으며 아래의 DTD를 데이터베이스 테이블에 저장한다.

```
<!ELEMENT paper
     (paper-title,author,document) >

<!ELEMENT paper-title (#PCDATA) >

<!ELEMENT author (#PCDATA) >

<!ELEMENT document
                (abstract, paper-body) >

<!ELEMENT abstract (#PCDATA)  >

<!ELEMENT paper-body (#PCDATA) >
```

표 1 DTD 테이블

3.3 XML문서의 DB전환

XML 문서는 구조적인 문서이다. 그러므로 XML 문서는 문서를 개별적으로 가지고 있는 것보다 데이터베이스에 저장하여 관리하는 것이 보다 효과적이다. 이렇게 하는 것이 문서의 검색과 관리를 훨씬 쉽게 해주기 때문이다[6]. 그러므로 XML 문서를 태그별로 구분하는 것이 필요하며, 구분된 태그별로 문서를 저장하는 저장 기법과 반대로 문서를 다시 구성하는 재구성 기법도 필요하다.

▶ DB 저장기법

1. 각각의 태그는 그에 해당하는 각 테이블에 저장된다.
2. 하나의 문서에 해당하는 서지정보의 키 값을 저장한다.
3. 그 문서에 포함되는 본문을 태그별로 구분한다.
4. 서지정보와 함께 태그의 내용을 저장한다.
5. 본문에 포함되는 태그의 순서를 함께 저장한다.

3.4 저장 구조

구조문서의 트리 구조를 저장하기 위하여 트리를 각각 노드로 작게 나누어 데이터베이스에 저장하고 관리할 경우에, 어떤 특정 서브트리 이하는 구조적인 정보의 저장이 필요 없다고 가정하면 저장되는 구조정보는 사용되지 않으면서 관리되어져야 하는 불필요한 정보가 될 수 있다. 이렇게 불필요한 부분은 통합되어서 하나의 커다란 노드처럼 트리를 관리하는 방법이 효율적이다.

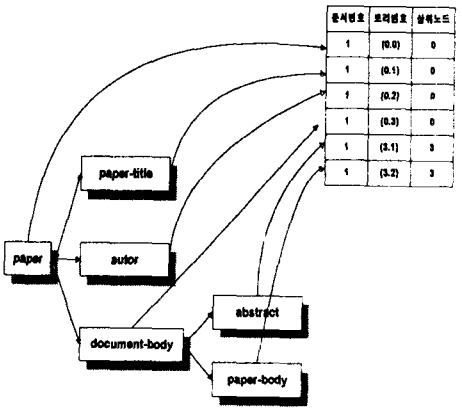

그림 2 문서 저장 구조

이러한 트리 구조의 저장단위를 지정하는 방법

은 저장된 문서를 복구할 때 매우 효과적으로 복구가 가능하다. 그리고 저장할 경우 불필요한 정보의 저장을 최대한 줄일 수 있다는 장점이 있다.

그림 2은 XML 트리구조를 데이터베이스 테이블에 저장하기 위하여 각각의 트리를 매치시켜 저장한다.

3.5 문서 검색 시스템

기존의 전문 검색 시스템들은 문서에 색인어를 부여함으로써 문서검색을 가능케 하였다. 문서로부터 수동 혹은 자동으로 색인어를 추출하여 사용자의 질의가 주어졌을 때, 질의에 사용된 단어와 문서 색인어 사이의 유사성을 계산하여 결과를 제공하였다. 이러한 시스템들은 문서를 단순히 단어의 집합으로 보며 문서의 구조 정보는 무시한다. 그렇지만 XML 문서는 구조적 정보를 담고 있는 엘리먼트로 구성되어 있기 때문에 엘리먼트를 기반으로 한 엘리먼트 구조에 대한 검색이 필요하다. 본 시스템에서 검색의 주된 데이터는 사용 DTD 정의에 의해 작성된 테그를 바탕으로 작성된 XML 문서를 대상으로 한다. 즉 문서에 기술한 문서요소를 기반으로 제목, 저자, 본문, 초록 등으로 데이터베이스를 구축하고 검색을 실시한다.

4. 결론

인터넷등 기술발전에 따른 급속한 사무환경의 온라인 화에 따라, Semi-structured 데이터가 양산되고 있다, Semi-structured 데이터는 비정형, 불명확한 구조, 데이터와 스키마가 혼재되어 있다.

기존의 문서관리 시스템의 제한점인 포맷에 따른 문서의 내용, 구조, 표현의 분리 문제, 메타정보의 자동화 문제, 문서단위의 검색 문제는, XML을 사용하여 해결할 수 있다.

이 모델은 내용 및 구조 검색의 경우 Element에 대한 속성지원, 부분문서의 순서 유지 등의 기존 데이터 모델의 제약점을 해결하고 있다. 또 문서의 구조에 대한 질의나 내용과 구조가 통합된 질의가 가능하며 검색단위를 문서에서 Element로 세분화하였다.

참고문헌

[1] T. Bray et al "Extensible Markup Language (XML)1.0", "http://www.w3.org/TR/REC-xml-19980210

[2] Jon Bosak, 1997.3, "XML, Java, and the Future of the Web", http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm

[3] Richard Light. Presenting XML. SAMS NET.

[4] 이용배, 손기락 "SGML 문서 저장을 위한 스키마 생성기 및 자동삽입기의 설계 및 구현", 정보과학회 가을 학술발표 논문집 1997.

[5J. Robie et al, "XML Query Language(XQL)," http://www.w3.org/TR/1998/NOTE-xml-ql-19980819/

[6] D. Florescu and D.Kossmann, "Storing and Querying XML Data Using and RDBMS," Bulletin of the Technical Committee on Data Engineering, Vol.22, No. 3, 1999, pp. 27-34.