구조 정보 검색을 위한 XML 저장관리시스템 설계 및 구현

이종설*, 박종관*, 정연수*, 손충범*, 강형일*, 유재수*, 김동율**, 최한석***
*충북대학교 정보통신공학과
**(주) 한국지식웨어
***목포대학교 정보공학부

Design and Implementation of an XML Repository System for Structural Retrieval

Jon Sul Lee*, Chong Kwan Park*, Yon Soo Jong*, Chung Beom Sohn*, Hyung Il Kan*,
Jae Soo Yoo*, Dong Yul Kim**, Han Suk Choi***
Dept. of Computer & Communication Eng. Chungbuk National University

요 약

본 논문에서는 대용량의 XML 문서를 효과적으로 저장, 관리 및 구조 기반 검색이 가능한 XML 저장관리시스템을 설계하고 구현한다. 구현한 XML 저장관리시스템은 관계형 모델을 기반으로 하고, XML 문서 전체를 저장하는 비분할 저장 모델을 사용하며, DTD에 따라 스키마가 생성되는 동적 스키마 생성 모델을 특징으로 한다. 본 논문의 XML 저장관리시스템은 BRS 검색엔진과 ORACLE을 기반으로 하며 질의처리기 및 검색결과생성기, XML 객체관리자, XML 인덱스관리자, 구조검색엔진 등으로 구성된다. 이를 통하여 내용 및 애트리뷰트 검색 뿐만 아니라 다양한 구조 정보검색을 효율적으로 지원한다.

1. 서 론

차세대 웹 문서 포맷으로 부각하는 XML(eXtensible Markup Language)은 W3C(World Wide Web Consortium)에서 제안된 국제 표준의 전자문서 메타 언어이다[1]. XML은 웹에서 구조화된 문서를 전송 가능하도록 설계된 표준화된 텍스트 형식으로, 문서를 구성하는 각 요소들의 독립성을 보장하게 함으로서 문서의 호환성, 내용의 독립성, 요소 변경의 용이성 등의 특성을 제공한다. 이에 따라 HTML의 새로운 대안으로 떠오르고 있는 XML이 인터넷 관련업계의 주요 관심사가 되고 있다. 현재 인터넷 Web 문서뿐만 아니라 전자도서관, CSCW(Computer Supported Cooperative Work) 그리고 CALS(Commerce At the Light Speed)를 포함한 다양한 분야에서 XML을 활용하고자 폭 넓은 연구를 하고 있으며, 수학 분야의 MathML, 채널 기술의 CDF(Channel Definition Format), 이동통신에서의 HDML(Handheld Device Markup Language) 등 최근에 구체화되는 응용이 부쩍 많아지고 있다[3].

이와 같이 전세계적으로 XML에 대한 관심이 고조되고 실제 많은 분야에서 활용되고 있기 때문에 향후 XML 문서의 확산이 더욱 가속화될 것이며, XML 문서의 생명주기 전 과정에서 정보의 생산성, 재사용성, 지속성, 이식성 등과 같은 XML 문서 사용의 장점들을 얻기 위해서는 XML 문서들을 저장, 관리 및 검색할 수 있는 XML 저장관리시스템이 필수적으로 요구된다[2,3,5].

이에 본 논문에서는 XML 문서의 구조정보를 이용하여 문서의 정보를 효율적으로 저장, 관리 및 XML 인스턴스에 대한 내용검색 뿐만 아니라 구조검색을 지원하는 XML 저장관리 시스템을 설계하고 구현한다. 이를 위해 현재 가장 많이 쓰이고 있는 관계형 데이터베이스인 오라클을 저장시스템으로 활용하고 인덱스 생성 및 내용기반 검색을 위해 BRS 검색엔진을 이용한다. 관계형 데이터모델은 현재 널리 사용되는 관계형 데이터베이스를 사용함으로써 데이터베이스의 활용이 쉬우며 상용 검색엔진인 BRS를 사용함으로써 데이터베이스의 키워드 검색능력보다 탁월한 검색능력을 보일수 있다.

본 논문의 구성은 다음과 같다. 2장에서 XML을 중심으로 기존 관련 연구를 살펴보고 3장에서는 XML 설계한 저장관리시스템을 설명한다. 4장에서는 구현한 XML 저장관리시스템의 구조 및 각 모듈에 대해 알아본다. 5장에서는 마지막으로 결론 및 향후 연구 방향을 제시한다.

2. 관련연구

XML 저장관리시스템에서 가장 핵심적인 역할을 수행하는 XML 객체 관리자를 개발함에 있어서 가장 먼저 선행되어야 하는 것이 데이터 모델링이다. XML 문서 모델링에 대해서 기존에 연구되었던 내용들에 대해서 살펴보면 DBMS의 활용에 따른 분류에는 관계형 모델과 객체지향 모델로 나누어진다[2,6].

관계형 모델은 현재 가장 많이 사용하고 있는 관계형 데이터베이스를 기반으로 하고 있기 때문에 쉽게 접근할 수가 있어 사용자들의 전반적인 확산이 빠르다. 그러나, 관계형 데이터베이스의 특성상 문서의 구조에 대한 충분한 정보를 유지하기 위해 필요로 하는 테이블과 튜플의 수가 기하급수적으로 늘어날 수 밖에 없으며, 이에 따른 JOIN 연산으로 인한 시스템의 성능이 저하되는 단점이 있다. 객체지향 모델은 데이터베이스에서 지원하는 객체지향 개념을 이용할 수 있기 때문에 상속과 같은 객체지향 특성을 이용할 수 있으며, 엘리먼트 간의 전후종속 관계를 클래스에 기반한 객체들간의 링크로 나타낼 수 있기 때문에 구조적인 문서를 모델링 하는데 적합하다 할 수 있다.

저장방식에 따른 분류에는 분할 저장 모델과 비분할 저장 모델, 그리고 혼합모델로 나누어진다. 분할 저장 모델은 XML 인스턴스를 엘리먼트 별로 나누어서 저장한다. 이 모델은 문서의 일부 내용들이 수정되었을 때 관계되는 노드들만 수정하면 되므로 문서의 편집 및 관리가 쉽고, 동일한 내용을 갖는 노드들을 공유할 수 있다는 장점이 있지만, 문서의 내용을 추출하고자 할 때 각 단말 노드들을 순회하며 통합하는 과정에서 시스템의 성능을 저하시키는 문제가 발생한다. 비분할 저장 모델은 XML 문서 전체를 BLOB 형태로 저장한 다음, 각각의 단말 노드는 오프셋 정보를 가지고 접근하는 방식이다. 이는 문서를 한꺼번에 저장하였기 때문에 통합 과정이 필요 없어 문서 참조를 빨리 할 수 있지만, 내용의 일부만이 수정되었을 때도 문서 전체를 재구성해야한다는 큰 단점이 있다. 혼합모델은 분할 저장 모델과 비분할 저장 모델을 혼용하여 사용하는 모델로 각각의 모델에서 단점을 보완하고자 상대 모델의 특성을 일부 포함하였다. 하지만 혼합모델의 단점인 저장 공간이 많이 소모된다는 문제점이 있다.

3. XML 저장관리시스템 설계

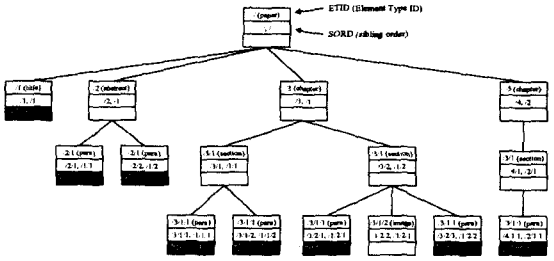
일반적인 저장관리시스템에 비해 XML 저장관리시스템은 XML이

갖고 있는 다음과 같은 특성을 고려하여야 한다[4,5,7]. XML 문서는 복잡한 구조와 다양한 미디어를 포함할 수 있다. 이러한 XML 문서는 구조정보를 이용하여 문서를 효율적으로 관리하기 때문에 데이터베이스에 XML 문서, DTD, 구조정보 및 다양한 미디어를 저장, 관리해야 된다. 또한 XML 문서의 특성을 볼 때 기존의 정보검색시스템(IRS: Information Retrieval System)이 제공하지 못한 문서의 논리적인 계층 구조를 이용한 검색, 엘리먼트가 갖는 속성에 대한 검색 등을 지원하여야 한다. 이와 같은 XML 문서의 특성을 반영하기 위해 다음과 같이 구조정보 표현 및 스키마를 설계하였다.

3.1 구조정보 표현

구조정보는 색인정보나 데이터베이스의 접근을 최대한 줄이는 방안으로 객체의 ID에 문서의 구조 정보를 갖도록 하는 방법을 주로 사용하였다. 기존의 구조정보는 특정 엘리먼트 검색 시 ID 자체로는 검색이 어렵게 된다. 이에 본 연구에서는 부모, 자식, 형제 등 문서 트리구조의 특정 엘리먼트 및 특정 깊이에 대한 질의가 가능한 구조정보표현을 제안한다.

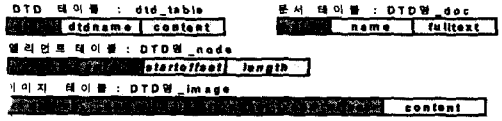
이를 위해 XML 문서의 구조정보를 표현하기 위해 ETID(element type ID), SORD(sibling order), SSORD(same sibling order)를 이용하였다. ETID(element type ID)는 XML문서의 논리적 구조를 이루는 DTD의 각 엘리먼트에 할당되는 유일한 값으로 문서 구조를 트리 형태로 표현하여, UNIX 파일 시스템에서 디렉토리를 표현하는 방법과 같은 방법을 이용하여 부여한다. 엘리먼트간의 형제 노드들간의 순서정보(SORD : sibling order)와 동일 타입의 엘리먼트들 간의 순서정보(SSORD : same sibling order)의 표현은 XML 문서 인스턴스에 적용한다. 형제 노드들간의 순서정보(SORD)는 동일 부모를 갖는 엘리먼트들의 출현 순서이며, 동일 부모를 갖는 엘리먼트들 중 동일한 형(type)간의 순서는 SSORD로 표현된다. 자식 엘리먼트의 순서정보에는 부모 엘리먼트의 순서정보도 함께 표현되는데 표현방법은 ETID의 표현방법과 동일하다. 다음 (그림 1)는 XML 문서의 구조정보를 본 연구에서 제안하는 방법으로 트리 형태로 표현한 예이다.



(그림 1) XML 문서의 구조 정보

3.2 스키마 구조

XML 문서는 복잡한 구조와 다양한 미디어를 포함하기 때문에 데이터베이스에 XML 문서, DTD, 구조정보 및 다양한 미디어 등을 저장, 관리해야 된다. 이를 위해 XML 데이터 모델링이 필요하며 본 논문에서는 하부 저장 시스템으로 사용하는 ORACLE 기반의 관계형 모델을 사용하였으며, XML 문서를 저장함에 있어서 문서의 빠른 추출을 위하여 비분할 저장 모델을 고려하였다. 본 논문에서 구현한 XML 저장관리시스템을 위한 스키마 구조는 (그림 2)과 같다.



(그림 2) 스키마 구조

DTD 테이블은 저장할 XML 문서의 DTD를 저장, 관리하며 문서 테이블은 XML 문서 전체를 저장, 관리한다. 엘리먼트 테이블은 구조정보 추출기를 통해 추출된 XML 문서의 각 엘리먼트의 구조정보를 저장, 관리한다. 또한 XML 문서에 이미지를 포함하는 경우 이미지 저장이 필요하다. 이미지 테이블은 이미지가 정의된 애트리뷰트의 문서번호, 이미지 파일명, XML 문서 안에서의 시작위치와 끝나는 위치 정보를 저장한다.

3.3 질의 분석

XML 문서에 대한 검색은 크게 내용검색, 구조검색, 애트리뷰트검색 그리고 내용+구조 등의 혼합 검색으로 나눌 수 있다.
다음은 본 논문에서 설계한 시스템에서 지원되는 질의 형태와 질의 예를 보여준다.

● 내용 질의
예) "XML"이라는 단어를 포함하는 문서를 찾으시오.

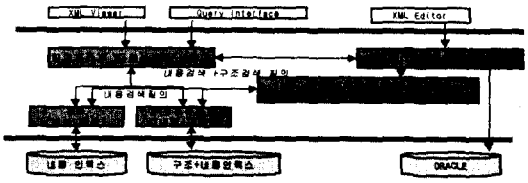
● 구조 질의
1. 특정 엘리먼트 질의
예) "XML"이라는 단어를 포함하는 첫 번째 Chapter를 찾아라.
2. 특정 엘리먼트의 부모, 자식, 형제 엘리먼트를 질의
예) "XML"이라는 단어를 포함하는 첫 번째 Chapter의 부모 엘리먼트를 찾아라.
예) "XML"이라는 단어를 포함하는 Chapter의 두 번째 자손 엘리먼트를 찾아라.

● 혼합 질의
1. 특정 키워드를 갖는 엘리먼트의 부모, 자식, 형제 엘리먼트를 질의
예) "XML"이라는 단어를 갖는 Chapter의 자손 엘리먼트인 Section들 중 "SGML"이라는 단어를 갖는 것을 찾아라.

● 애트리뷰트 질의
1. 엘리먼트에 나타날 수 있는 애트리뷰트 이름과 특성값에 대한 질의
예) 애트리뷰트 "check" 가 "yes"인 엘리먼트를 찾아라.

4. XML 저장관리시스템 구현

본 논문에서 구현한 XML 저장관리시스템은 하부 저장시스템으로 ORACLE 7.3을 사용하였으며, ORACLE에 비해 일반적인 키워드 검색이 뛰어나며, 길이가 정해지지 않는 문자열에 대한 인덱스 생성을 위해 BRS 검색엔진을 활용하였다. (그림 3)은 구현한 XML 저장관리시스템 구조를 보여 준다. XML 저장관리시스템은 질의처리 및 검색결과 생성기, XML 객체 관리기, XML 인덱스 관리기, BRS 검색엔진, 구조검색엔진으로 구성된다. XML 저장관리시스템을 구성하는 각 모듈의 역할은 다음과 같다.



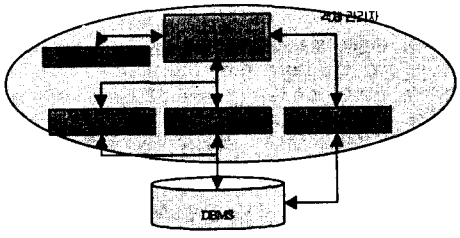
(그림 3) XML 저장관리기 시스템 구성도

XML 저장관리시스템에서 가장 핵심적인 기능을 담당하는 XML 객체관리자에서는 실제 XML 문서를 저장하기 위한 스키마 생성 및 XML 문서 인스턴스의 저장 및 추출을 담당한다. XML 인덱스관리자는 구조검색, 애트리뷰트검색을 처리하는 인덱스를 생성 및 관리한다. BRS 검색엔진은 내용인덱스를 생성, 관리하며 사용자의 질의 중 키워드 검색을 처리한다. 구조검색엔진은 BRS에서 처리하지 못하는 구조검색, 애트리뷰트검색, 혼합검색을 지원하기 위해 자체 구현하였다. 질의처리기에서는 사용자 질의를 분석하여 내용검색은 BRS 검색엔진이 처리하도록 하고, 구조검색, 애트리뷰트검색, 내용+구조검색 등 혼합검색은 구조검색엔진을 사용한다. 또한 검색결과생성기에서는 BRS와 구조검색엔진이 처리한 검색 결과를 이용하여 문서 전체 또는 일부분을 사용자에게 제공한다.

4.1 XML 객체관리자

XML 객체관리자는 XML 문서를 효율적으로 관리하기 위한 문서의 일관된 관리 기능을 제공한다. 이를 위해 먼저 XML 객체관리자는 XML 문서의 데이터 모델링에 의해 XML이 포함하고 있는 여러 가지 특성들을 고려하여 스키마를 설계하고, 데이터베이스에 생성한다. 또한 실제 XML 문서, 구조정보, 이미지 등을 데이터베이스에 저장하며, 저장된 XML 문서를 사용자가 원하는 전체 문서 또는 문서 일부분을 꺼내는 일을 담당한다.XML 객체관리자를 구성하는 세부 모듈은 (그림 4)와 같이 객체 저장관리기, 구조정보 추출기, XML 인스턴스 관리기, XML 인스턴스 저장기, 스키마 생성기 등으로 이루어진다. 객체 저장관리기는 XML 객체관리자를 구성하는 각 모듈들에 대한 통합 인터페이스를 제공하며, 구조정보 추출기는 XML 인스턴스에서 문서를 저장하는데 필요한 구조 정보를 추출하는 역할을 한다. XML 인스턴스 저장기는 XML 문서, 구조정보 추출기에서 추출된 문서의 구조정보, 이미지 등을 데이터베이스에 저장하기 위한 모듈이다. XML 인스턴스 관리기는 사용자가 요구하는 문서 전체 혹은 문서 일부분과 이미지를 데이터베이스로부터 추출하는 기능을 담당한다. XML 스키마 생성기는 DTD 테이블을 시스템 초기화에 생성하

고 이후 다양한 DTD를 수용할 때마다 문서테이블, 엘리먼트 테이블, 이미지 테이블을 동적으로 생성한다.



(그림 4) 객체관리자의 구성도

4.2 XML 인덱스관리자

XML 문서에 대해 내용, 구조, 애트리뷰트, 혼합검색을 지원하기 위한 색인구조를 만들고 색인 정보를 관리하는 모듈이 XML 인덱스 관리자이다. XML 인덱스 관리자에서는 full-text 검색을 위한 내용 색인기, 구조 검색과 구조와 내용을 동시에 지원하는 XML 구조+내용 색인기, 애트리뷰트 검색을 지원하기 위한 애트리뷰트 색인기로 구성된다. 내용 색인기에서는 빠른 full-text 검색을 위해 BRS 검색엔진의 색인을 이용하고 검색 결과로서 실제 문서는 BRS 검색엔진에서 부여한 식별자와 객체관리자에서 부여한 식별자 정보를 유지하는 매핑 정보를 이용하여 사용자에게 보내진다.

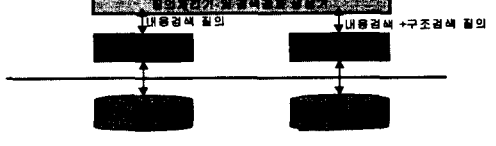
(그림 5)는 본 논문에서 설계한 구조+내용 색인 구조이다. 그림에서 보듯이 구조질의와 내용이 혼합된 질의를 빠르게 처리하기 위해 인덱스 화일에서는 엘리먼트 식별자인 ETID와 키워드를 가지고 인덱싱한다. 이렇게 함으로서 엘리먼트 기반의 내용 검색을 빠르게 지원할 수 있다. 포스팅 화일에서는 검색 결과로서 해당하는 특정한 엘리먼트나 문서를 접근하기 위한 정보들로 구성된다. 특히 이 정보들 중에 SORD와 SSORD를 이용하여 여러 XML 문서에서 고유한 엘리먼트를 접근할 수 있다. (그림 6)은 애트리뷰트 검색을 지원하기 위한 색인 구조이다.

(그림 5) 구조+내용 색인구조

(그림 6) 애트리뷰트색인 구조

4.3 XML 질의 처리기

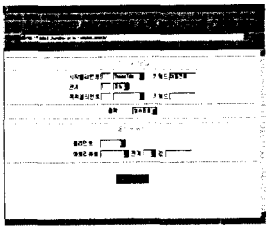
질의처리 및 검색결과 생성기는 내용검색과 구조검색을 구분하여 내용검색이 사용자로부터 요청될 경우는 BRS 검색엔진에서 처리하고 처리된 결과는 검색 결과생성기에서 XML 객체관리자의 XML 인스턴스관리기를 통해 ORACLE에 저장되어 있을 문서전체 혹은 일부분을사용자에게 보여 준다. 구조검색, 애트리뷰트검색, 혼합검색이 요청될 경우는 구조 검색 엔진이 검색한 결과를 이용하여 문서 전체 또는 일부분을 사용자에게 제공한다. (그림 7)은 질의 처리의 과정을 보여준다.



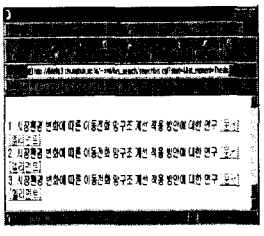
(그림 7) 질의 처리 블럭도

본 논문에서는 구현한 시스템을 테스트하기 위해 XML문서에 대한 다양한 검색이 가능하도록 웹기반의 질의 인터페이스를 설계하였다. 내용 검색은 XML에 표현되어 있는 텍스트에 대한 스트링 매칭 검색을 하고, 구조 검색에서는 계층간의 관계를 위해 조상, 자손의 관계가 있으며, 같은 계층내의 관계를 위해 형제 관계가 있다. 또한 각 엘리먼트들간의 순서에 따라 선후 관계(+ , - , 숫자)를 갖을 수 있도록 하였다. 애트리뷰트 검색은 애트리뷰트 값의 다양한 관계 연산을 제공한다. 검색 결과로는 문서 전체 또는 엘리먼트를 지정할 수 있도록 하였다.

(그림 8), (그림 9)는 「 "이동전화"를 포함하는 ThesisTitle의 두번째 뒤(following) 형제 엘리먼트를 찾아라 」 라는 구조 + 내용 질의 처리 예를 보여준다.



(그림 8) 구조+내용 질의 인터페이스



(그림 9) 구조+내용 결과

5.결론

본 논문에서 설계 및 구현된 대용량의 XML 저장관리시스템은 관계형 데이터베이스를 사용하는 관계형 모델과, 추출 성능향상을 위해 XML 문서의 전체를 저장하는 비분할 모델, 그리고, 모든 DTD 문서를 수용할 수 있는 형태의 동적인 스키마 생성을 특징으로 한다. XML 저장관리시스템은 질의처리기 및 검색결과생성기, XML 객체관리자, XML 인덱스관리자, 구조검색엔진 등으로 구성된다

XML 문서의 효율적인 관리를 위하여 XML 객체관리자에서는 구조정보추출기, 객체 저장관리기, 스키마 생성기, XML 인스턴스 저장기, XML 인스턴스 관리기 등 세부 모듈을 구현하였다. XML 인덱스 관리자는 구조검색과 내용검색을 동시에 지원하기 위한 XML 구조+내용 색인기, 그리고 애트리뷰트 검색을 위한 XML 애트리뷰트 색인기로 구성하였다. 질의처리 및 검색결과 생성기는 내용검색과 구조검색을 구분하여 사용자로부터 내용검색이 요청될 경우는 BRS 검색엔진에서 처리하고 구조검색 질의와 내용검색 질의가 혼합되어 요청이 될 경우는 구조검색엔진에서 처리하며 결과를 이용하여 검색결과생성기에서 문서 전체 또는 일부분을 사용자에게 제공한다. 또한 상용 검색엔진인 BRS에서 수행하는 내용검색 이외의 구조검색, 애트리뷰트검색을 처리할 수 있는 구조검색엔진을 개발하였다. 마지막으로 구현한 구조검색엔진의 성능을 테스트하기 위해 검색 인터페이스를 구현하여 테스트하였다.

향후연구과제로는 XML 문서가 가지고 있는 구조정보는 일반적으로 트리 형태의 계층적 구조로 표현 될 수 있기 때문에 객체지향 데이터베이스로 관리하는 것이 효율적이다. 이에 객체지향형 모델링을 통한 XML 저장관리시스템 설계 및 구현을 하고자 한다.

참고 문헌

[1] 손정한, 이희주, 장재우, 심부성, 주종철, "구조화된 문서를 위한 정보검색시스템의 설계 및 구현", '98 동계 데이터베이스 학술대회 논문집 제14권 1호, PP102-106, 1998

[2] 연제원, 장동준, 김용훈, 이강찬, 이규철, "효율적인 검색 지원 SGML 저장 관리기의 설계 및 구현", '99 한국 데이터베이스 학술대회 논문집 15권 1호, pp136-143, 1999

[3] 유재수외 8명, "전자도서관 표준문서관리를 위한 XML 저장관리기 기술 개발", 케이오텍 최종보고서,1999

[4] Charles L. A. Clarke, Gordon V. Cormack, Forbes J. Burkowski, An Algebra for Structured Text Search and a Framework for its Implementation, The Computer Journal 38(1), pp. 43-56, 1995.

[5] Francois, "Generalized SGML repositories: Requirements and modelling", Computer Standards & Interfaces, 1996

[6] Brian Lowe, Justin Zobel, Ron Sacks-Davis, A Formal Model for Databases of Structured Text, DASFAA 1995, pp. 449-456.

[7] Ian A. Macleod, Storage and Retrieval of Structured Documents, Information Processing and Management, vol. 2