XML 문서의 효율적인 저장구조와 색인 모델의 설계

김은정
부산 외국어대학교 컴퓨터전자공학부
ejkim@taejo.pufs.ac.kr

Design of Efficient Storage Structure and Indexing Model of XML Document

EunJung Kim
Dept. of Computer Engineering, Pusan University of Foreign Studies

요 약

XML 문서는 문서의 내용뿐 아니라, 의미를 가지는 구조 정보, 그리고 다양한 의미를 부과할 수 있는 링크 정보를 가지고 있다. 본 논문에서는 XML 문서를 보다 효율적으로 관리하기 위하여 DTD와 XML 문서에 대한 새로운 저장 방법과 이를 이용한 색인 모델을 제안한다. 이를 위해 하나의 XML 문서를 저장함에 있어, 엘리먼트 구조 정보, 애트리뷰트 정보, 링크 정보의 구성 방법을 제시하고, 이를 바탕으로 링크 정보를 이용한 내용 검색 색인 모델과 구조 검색, 애트리뷰트 검색을 위한 색인 모델을 설계한다. 또한 제안된 모델에서의 사용자들의 다양한 질의 유형의 처리 과정을 설명한다.

1. 서 론

웹의 눈부신 발달과 사용자들이 요구하는 정보의 양이 급증하고 또한 사용자들간에 정보를 서로 공유하고 교환하는 응용 범위 또한 매우 광범위해져 가고 있다. 이에 기존의 HTML이 문서의 내용을 화면에 보여주는 것에 중점을 두고 있기 때문에, 서로간에 의미를 가진 문서의 공유나 응용에는 한계가 있다. 이에 대한 보완으로 특정 응용 분야에 사용될 수 있는 문서의 내용에 의미를 부여하여 사용자간에 자료를 보다 효율적으로 공유하고 교환할 수 있는 구조적 문서로서 XML이 각광을 받고 있다.

DTD를 기반으로 한 XML 문서는 의미를 가지는 구조 정보와 문서와 문서간 다양한 의미를 부여할 수 있는 링크 정보, 그리고 다양한 종류의 애트리뷰트 정보를 가지고 있다. 이 중에서 구조 정보는 문서가 내포하는 정보 관리를 보다 효율적으로 할 수 있으며, 의미를 부여한 검색을 효율적으로 수행하는데 이용할 수 있다. 구조적 문서에 대한 검색은 문서의 내용에 대한 검색뿐만 아니라, 의미를 가진 구조에 대한 검색과 구조와 내용이 혼합된 검색 등 사용자의 다양한 질의에 대한 검색이 가능해야 한다. 따라서 이를 기반으로 구조기반 질의를 지원하는 구조적 검색에 대한 연구가 많이 진행되어 왔다[1,2,4]. 이러한 기존의 연구에서는 문서가 가지는 구조 정보만을 다루기 때문에 문서와 문서, 또는 문서와 문서의 특정 엘리먼트 사이의 다양한 의미를 내포하는 링크 정보는 제외된다.

XML 링크에는 다양한 의미를 가지고 있어서 링크된 문서 사이의 관계를 다양하게 정의할 수 있다. 따라서 이러한 링크 정보를 검색에 활용하는 연구가 많이 진행되어 왔다. 이전의 연구[3]에서 이미 XML 링크 정보를 활용한 검색 시스템을 설계한 바 있다. 본 논문에서는 XML 문서의 구조 정보와 링크 정보를 보다 효율적으로 관리할 수 있는 저장 구조를 제시하고, 사용자들의 다양한 질의를 처리할 수 있는 색인 모델을 설계한다. 저장 구조에서는 먼저 DTD의 구조 정보를 저장하고, 이를 기반으로 한 하나의 XML 문서의 저장 구조는 엘리먼트 구조 정보, 애트리뷰트 정보, 링크 정보로 구성된다. 다음으로 DTD와 XML 문서의 저장 구조를 바탕으로 사용자들의 내용 질의, 구조 질의, 혼합 질의 등 다양한 질의를 지원할 수 있는 색인 모델을 설계한다. 설계된 색인 모델 중 내용 색인 모델에서는 문서의 구조 정보와 링크 정보를 함께 사용하여 설계한다.

2장에서는 DTD와 XML문서의 구조 정보와 링크 정보의 구성 방법을 제시하고, 3장에서는 제시된 구조 정보를 바탕으로 색인 모델을 설계한다. 4장에서는 여러 가지 유형의 질의어 처리 과정을 설명하고 마지막으로 5장에서 결론을 보인다.

2. DTD와 XML 문서 정보 구성 방법

2.1 DTD 구조 정보

DTD 구조 정보에는 엘리먼트 이름, 각 엘리먼트 이름을 구별하기 위한 식별자(ID), 엘리먼트의 상위 엘리먼트, 그리고 하위 엘리먼트로 구성된다. 엘리먼트 이름은 DTD에 나오는 모든 엘리먼트를 말하며, 식별자(ID)는 각 엘리먼트를 구별하기 위한 유일한 값으로 부모-자식 엘리먼트의 순서와 상관없이 나열한 순서대로 10진수를 부여한다. 상위 엘리먼트는 각 엘리먼트의 부모 엘리먼트에 대한 ID를 순서대로 나열하고, 하위 엘리먼트는 자식 엘리먼트의 ID를 순서대로 나열한다. DTD 구조 정보를 구성함에 있어서는 애트리뷰트와 링크 정보, 발생 횟수에 대한 정보는 포함하지 않는다. 이러한 정보는 XML 문서 정보를 구성할 때 저장한다. (그림 1)에서 간단한 DTD의 내용을 보이고 (그림 2)에서 (그림 1)의 DTD에 대한 구조 정보를 저장하기 위한 DTD 구조 정보 테이블을 보인다.

```
<! ELEMENT paper              (head, body, reference+)>
<! ELEMENT head               (koreanhead , englishhead )>
<! ELEMENT   koreanhead       (koreantitle , koreanauthor +, koreanabstract )>
<! ELEMENT   koreantitle      (#PCDATA)>
<! ELEMENT   koreanauthor     (#PCDATA)>
<! ATTLIST   koreanauthor
               xmlns :xlink   CDATA #FIXED   "http://www.w3.org/1999/ xlink"
               xlink :type    CDATA #FIXED   "simple "
               xlink :href    CDATA #REQUIRED
               xlink :show  (new|replace|embed)     "replace "
               xlink :actuate ( onRequest |onLoad ) "onRequest "
>
<! ELEMENT   koreanabstract   (#PCDATA)>
<! ELEMENT   englishhead      (englishtitle , englishauthor +, englishabstract )>
<! ELEMENT   englishtitle     (#PCDATA)>
<! ELEMENT   englishauthor    (#PCDATA)>
<! ELEMENT   englishabstract  (#PCDATA)>
<! ELEMENT body               (chapter+)>
<! ELEMENT chapter            (#PCDATA | section+)>
<! ELEMENT section            (#PCDATA |      para )* >
<! ELEMENT   para             (#PCDATA)>
<! ELEMENT reference          (#PCDATA)>
<! ATTLIST   reference
               xmlns :xlink   CDATA #FIXED   "http://www.w3.org/1999/ xlink"
               xlink :type    CDATA #FIXED   "simple "
               xlink :href    CDATA #REQUIRED
               xlink :show  (new|replace|embed)     "replace "
               xlink :actuate ( onRequest |onLoad ) "onRequest "

>
```

그림 1 간단한 DTD 예

Element	ID	상위엘리먼트	하위엘리먼트
Paper	0	-	(1,10,14)
Head	1	(0)	(2,6)
Koreanhead	2	(0,1)	(3,4,5)
Koreantitle	3	(0,1,2)	-
Koreanauthor	4	(0,1,2)	-
Koreanabstract	5	(0,1,2)	-
Englishhead	6	(0,1)	(7,8,9)
Englishtitle	7	(0,1,6)	-
Englishauthor	8	(0,1,6)	-
Englishabstract	9	(0,1,6)	-
Body	10	(0)	(11)
Chapter	11	(0,10)	(12)
Section	12	(0,10,11)	(13)
Para	13	(0,10,11,12)	-
reference	14	(0)	-

그림 2 DTD 구조 정보 테이블

2.2 XML 문서 정보

하나의 XML 문서에는 문서 내용에 의미를 부여할 수 있는 구조 정보와 문서와 문서간 또는 하나의 문서안에서 엘리먼트간의 의미를 부여할 수 있는 링크 정보와 속성 정보를 가지고 있다. 이러한 모든 정보를 효율적으로 관리하기 위하여, 하나의 문서 인스턴스 정보를 구성함에 있어 엘리먼트 구조 정보, 링크 정보, 애트리뷰트 정보로서 구성한다(그림 3).

문서 인스턴스 정보

DOC_ID	ELE_INFO	ATTR_INFO	LINK_INFO

엘리먼트 구조 정보(ELE_INFO)

EID	DTD_ID	부모_EID	위치순서	자식수	내용

애트리뷰트 정보(ATTR_INFO)

EID	이름	값

링크 정보(LINK_INFO)

EID	링크_ID	HREF	ROLE

그림 3 XML 문서 정보 테이블

문서 인스턴스 정보는 문서를 구별하기 위한 식별자(DOC_ID), 문서안의 각 엘리먼트들간의 구조 정보를 저장하기 위한 엘리먼트 구조 정보(ELE_INFO), 엘리먼트가 가지는 애트리뷰트 정보(ATTR_INFO), 그리고 문서내 존재하는 링크 정보(LINK_INFO)로 구성한다.

엘리먼트 구조 정보는 EID, DTD_ID, 부모_EID, 위치순서, 자식수, 내용으로 구성된다. EID는 하나의 XML 문서내에 존재하는 모든 엘리먼트를 구별하기 위한 식별자로서, 문서에 나오는 모든 엘리먼트를 나열하고 부모_자식 엘리먼트와는 상관없이 10진수를 부여한다. DTD_ID는 각 EID가 DTD 구조 정보 테이블에서 가지고 있는 유일한 식별자이다. 따라서 하나의 문서에 존재하는 모든 엘리먼트들 중에서 같은 이름의 엘리먼트는 EID는 서로 다르지만, DTD_ID로서 해당 엘리먼트의 이름을 식별한다. 부모_EID는 각 엘리먼트가 문서상 어느 위치에 존재하는지를 식별하기 위해서 문서내 부모 엘리먼트의 EID를 순서대로 나열하여 저장한다. 위치순서는 (형제 엘리먼트들 사이의 위치, 동일한 이름의 형제 엘리먼트들 사이의 위치)로 구성된다. 자식수는 해당 엘리먼트의 하위 엘리먼트의 개수를 저장한다. 내용은 해당 엘리먼트가 텍스트를 가지는 엘리먼트일 경우, 텍스트의 내용을 저장한다. (그림 4)에서 (그림 1)의 DTD를 기반으로 하는 간단한 XML 문서를 보이고, (그림 5)는 문서의 엘리먼트 구조 정보 테이블이다.

```
<?xml  version= "1.0" encoding =  "enc-kr" ?>
<!DOCTYPE paper  SYSTEM        "paper. dtd" >
<paper>
   <head>
      <koreanhead  >
         <koreantitle  > 링크 의미를 이용한 하이퍼텍스트 ... </ koreantitle >
         <koreanauthor xlink  :href ="ejkim .html "> 김은정 </koreanauthor  >
         <koreanauthor  > 홍길동 </koreanauthor  >
         <koreanabstract  > 하이퍼텍스트 기반의 정보검색에 ... </koreanabstract  >
      </koreanhead  >
      <englishhead  >
         <englishtitle  > A model of hypertext retrieval     ... </ englishtitle >
         <englishauthor  > EunJung  Kim </ englishauthor  >
         <englishauthor  > Kil-Dong Hong </  englishauthor  >
      </ englishead  >
   </head>
   <body>
      <chapter> 1.  서론
         <section>
            <para > 서론 내용 ... </ para >
         </section>
      </chapter>
      <chapter> 2.  관련 연구
         <section>
            <para >  관련연구 내용 ... </ para >
         </section>
      </chapter>
      <chapter> 3.  연구내용 및 성능평가
         <section> 3.1  연구내용
            <para >  연구 내용 ... </ para >
         </section>
         <section> 3.2  성능평가  </ section>
      </chapter>
      <chapter> 4.  결론 및 향후과제  ... </ chapter>
   </body>
   <reference  xlink :href ="doc2.xml"> 참고 문헌 1 ... </ reference>
</paper>
```

그림 4 간단한 XML 문서

	EID	DTD_ID	부모_EID	위치순서	자식수	내용
Paper	0	0	/	(1,1)	3	-
Head	1	1	/0/	(1,1)	2	-
Koreanhead	2	2	/0/1/	(1,1)	3	
Koreantitle	3	3	/0/1/2/	(1,1)	0	
Koreanauthor	4	4	/0/1/2/	(2,1)	0	
koreanauthor	5	4	/0/1/2/	(3,2)	0	
Koreanabstract	6	5	/0/1/2/	(4,1)	0	
Englishhead	7	6	/0/1/	(2,1)	3	
Englishtitle	8	7	/0/1/7/	(1,1)	0	
Englishauthor	9	8	/0/1/7/	(2,1)	0	
Englishauthor	10	8	/0/1/7/	(3,2)	0	
Body	11	10	/0/	(2,1)	4	-
Chapter	12	11	/0/11/	(1,1)	1	-
Section	13	12	/0/11/12/	(1,1)	1	
Para	14	13	/0/11/13/	(1,1)	0	
Chapter	15	11	/0/11/	(2,2)	1	
Section	16	12	/0/11/15/	(1,1)	1	
Para	17	13	/0/11/15/16/	(1,1)	0	
Chapter	18	11	/0/11/	(3,3)	2	
Section	19	12	/0/11/18/	(1,1)	1	
Para	20	13	/0/11/18/19/	(1,1)	0	
Section	21	12	/0/11/18/	(2,2)	0	
Chapter	22	11	/0/11/	(4,4)	0	
Reference	23	14	/0/	(3,1)	0	

그림 5 엘리먼트 구조 정보 테이블

링크 정보 테이블은 EID, 링크_ID, HREF, ROLE로 구성된다. EID는 어느 엘리먼트에 속하는 링크인지를 식별하기 위하여 문서내 링크를 포함하는 엘리먼트의 EID를 저장한다. 링크_ID는 링크 식별자이다. 이전의 연구[3]에서 링크가 갖는 행동에 관련된 속성값에 따라 링크의 의미를 분류하여 링크 식별자를 정의하였다. 이 링크 식별자를 이용하여 링크_ID에 해당 링크를 분류하여 저장한다. HREF는 해당 링크가 지시하는 원격 문서 또는 원격 문서의 특정 엘리먼트의 주소이다. ROLE은 링크가 갖는 메타 데이터를 저장한다. 이 ROLE 값은 검색시 유용하게 활용할 수 있다. 본 논문에서는 ROLE 값을 활용하는 방법은 논하지 않는다. (그림 6)은 (그림 4)의 문서에 대한 링크 정보 테이블이다. 애트리뷰트 정보 테이블은 EID, 애트리뷰트 이름과 해당 속성 값을 저장하여 활용한다.

EID	링크_ID	HREF	ROLE
4	6	cjkim.html	-
23	6	doc2.xml	-

그림 6 링크 정보 테이블

3. 색인 구조

여기서는 2장에서 구성한 DTD 구조 정보 테이블과 XML 문서 정보 테이블을 기반으로 사용자의 다양한 유형의 질의를 효율적으로 처리할 수 있는 색인 구조를 설계한다. 이를 위해 키워드 색인 테이블, 구조 색인 테이블, 애트리뷰트 색인 테이블을 구성한다.

키워드 색인은 문서의 내용을 기반으로 한 검색을 처리하기 위하여 XML 문서에서 나타나는 색인어를 기준으로 색인 테이블을 구성한다. 구성은 색인 파일과 포스팅 파일로 이루어진다. 색인 파일의 구성 요소는 색인어, 색인어를 포함하고 있는 엘리먼트 개수이다. 포스팅 파일의 구성요소는 문서번호, EID, DTD_ID, 부모_EID, 위치순서, 지역빈도, 원격빈도이다. 문서번호는 해당 엘리먼트가 속해 있는 문서의 식별자이다. EID, DTD_ID, 부모_EID, 위치순서는 색인어를 포함하는 엘리먼트에 대한 구조 정보이다. 지역 빈도는 해당 엘리먼트에서 해당 색인어가 발생한 빈도수이다. 원격 빈도는 해당 엘리먼트가 링크를 포함하고 있을 경우에 링크가 지시하는 원격 문서에서의 색인어 발생 빈도이다. 원격 문서에서의 색인어 발생 빈도수를 계산하는 방법은 이전의 연구[3]를 바탕으로 한다(그림 7).

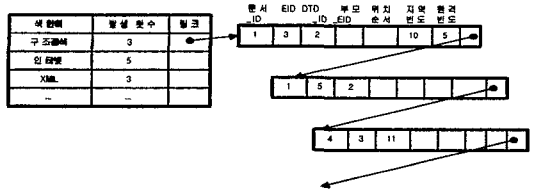

그림 7 키워드 색인 모델

구조 색인은 문서의 순수 구조를 기반으로 한 검색을 지원하기 위하여, XML 문서에서 텍스트를 포함하는 엘리먼트을 기준으로 색인한다. 색인 파일의 구성 요소는 엘리먼트 이름이고, 포스팅 파일은 문서 번호와 엘리먼트 구조 정보로 구성된다. 애트리뷰트 색인은 애트리뷰트 검색을 지원하기 위한 것으로, 색인 파일의 구성은 애트리뷰트 이름과 값으로 구성된다.

4. 질의처리 과정

이 장에서는 3장에서의 색인 모델을 기반으로 몇가지 유형의 질의어 처리 과정을 설명한다.

- 질의: "구조 검색"을 포함하는 문서를 검색하시오.
  처리 과정: 키워드 색인 테이블에서 "구조 검색"에 링크된 문서 번호와 단어 지역문서빈도와 원격 문서 빈도수를 이용해 문서를 나열한다.
- 질의 : <reference>가 있는 논문을 검색하시오.
  처리 과정 : 구조 색인 모델에서 <reference>를 검색하여 해당 문서를 나열한다.
- 질의 : <englishtitle>에 "structure retrieval"을 포함하는 논문을 검색하시오.
  처리 과정 :
  ① DTD구조정보Table의 <englishtitle>의 ID 선택.
  ② 키워드 색인 구조에서 "structure retrieval" 검색.
  ③ 검색된 엘리먼트 중에서 DTD_ID와 위의 ID가 동일한 문서를 찾아 나열한다.
- 질의 : 두 번째 chapter에 "구조 검색"을 포함하는 문서를 검색하시오.
  처리 과정 :
  ① DTD구조정보Table에서 <chapter>의 ID와 자식 ID를 가져온다.
  ② 키워드 색인 테이블에서 "구조 검색"을 검색한다.
  ③ 검색된 엘리먼트 정보와 위의 ID를 이용해 해당 문서를 선택한다.

5. 결론

본 논문에서는 XML 문서 기반의 검색 시, 보다 효율적인 문서의 관리와 이를 이용한 검색을 위하여 DTD와 XML 문서에 대한 새로운 저장 기법을 제시하고, 이를 기반으로 사용자들의 다양한 질의를 처리할 수 있는 색인 모델을 설계하였다. 앞으로는 XML 문서 집합을 대상으로 색인 모델의 효율성을 검증하고 아울러 기존 연구[3]와의 통합을 통하여 보다 효율적인 XML 문서 검색 시스템을 개발하는 것이다.

참 고 문 헌

[1] 박종관 외, "XML 문서의 효율적인 구조 검색을 위한 색인 모델", 한국정보처리학회 논문지, 제8-D권, p451~460, 2001.
[2] 김영자, 배종민, "GDIT를 기반으로 한 구조적 문서의 효율적 검색과 갱신을 위한 인덱스 설계", 한국정보처리학회 논문지, 제7권 2호, p411~425,2000.
[3] 김은정, 배종민, " XLinks를 이용한 하이퍼텍스트 검색 시스템", 한국정보처리학회 논문지, 제8권 5호, p483~494, 2001.
[4] D. W. Shin, .., "BUS:An Effective Indexing and Retrieval Scheme in Structured Documents", in Proc. Digitial Libraries, 1998.