XML 문서를 기반으로 한 Local 검색을 위한 색인 기법

정혜진°, 유춘식, 김용성
전북대학교 컴퓨터정보학과
hi-jin@hanmail.net, csyoo@cs.chonbuk.ac.kr, yskim@moak.chonbuk.ac.kr

A Indexed Technique for Local search based on XML Document

Hye-Jin Jeong°, Choun-Sick You, Yong-Sung Kim
Dept. of Computer Information, Chonbuk National University
hi-jin@hanmail.net, csyoo@cs.chonbuk.ac.kr, yskim@moak.chonbuk.ac.kr

요 약

오늘날 방대한 양의 정보를 관리하고 검색하기 위해 정보를 효율적으로 처리 할 수 있는 방안에 대해서 많은 연구가 진행되고 있다. 본 논문에서는 문서를 구조화하기 위한 방법으로써 XML을 기반으로 하고, 이를 효과적으로 검색하기 위해서 색인기법을 제안하므서 CD-ROM이나 하드 디스크와 Local 보조 기억 장치에 저장되어 있는 문서 파일을 효율적으로 검색할 수 있도록 한다.

1. 서론

오늘날 인터넷의 보편화로 정보 자료의 홍수상태로 인하여 자신이 원하는 정보를 얻기가 점점 어려워지고 복잡해지고 있어 보다 효과적으로 검색할 수 있는 연구가 활발히 진행되고 있다. 사용자는 자신에게 필요한 정보가 증가하면서 개인이 소유한 정보를 관리하기 위해서 CD-ROM이나 하드디스크와 같은 보조기억장치에 정보를 저장을 하고 있다.

그러나 이러한 보조기억장치에 존재하는 정보를 개인이 활용하기에는 여러 가지 문제점이 있다. 첫째, 정보가 파일 단위로 존재하기 때문에 원하는 정보를 빠른 시간 내에 찾기가 어렵고, 둘째, 정보를 조직적으로 관리하기가 힘들다.

이러한 문제점을 해결하기 위한 대안 중의 하나가 정보를 XML 문서의 형태로 관리하는 것이다. XML은 문서의 구조정보를 제공할 뿐만 아니라, XML 태그는 데이터를 해석하는데 사용할 수 있기 때문에 XML의 역할과 중요성이 인식되고 있다. 따라서 구조 정보를 내포하고 있는 데이터로서의 XML 문서를 효과적으로 관리하는 구조와 질의어 설계 및 처리에 대한 연구도 많이 진행되고 있다.

본 논문에서는 특정 엘리먼트에 대한 직접적인 접근이 가능하고 엘리먼트간의 관계를 구하는데 복잡한 연산이 필요하지 않도록 DTD에 나타난 엘리먼트들과 XML 문서의 구조 정보를 사용해서 XML 문서를 효율적으로 관리하고 검색할 수 있도록 정적 색인 모델 구조를 지양하고 보다 효율적인 검색을 할 수 있도록 새로운 색인 구조를 제안한다.[1].

2. 관련 연구

정보 환경의 변화는 사용자에게 쉽고 간편함을 제공하지만 데이터들은 과다하게 증가함으로써 복잡한 형태를 지닌다. 따라서 복잡한 정보를 처리하기 위한 노력으로 대용량 데이터베이스의 설계 및 자료의 저장구조 분석 방법 등의 연구로 이를 활용한 디지털 도서관, 전자출판, 전자 상거래 등의 다양한 분야에 활용되어 지고 있다. 또한 이러한 대량의 문서를 구조화하여 관리하거나 표준화된 방법으로 처리하기 위한 대안으로 XML 이 이용되어 지고 있다[5, 6].

그리고 DTD 정보를 기반으로 한 XML 기반의 Local 검색 시스템을 설계 구현하기 위해 XML소스 변환기로문서 파일을 XML 파일로 추가하는 연구도 있고[2]. 엘리먼트들 간의 계층 정보를 나타내는 ETID(Element Type ID), XML 문서에서 형제 엘리먼트의 순서정보인 SORD(Sibling ORDer), 같은 타입의 엘리먼트 형제에 대한 순서정보인 SSORD(Same Sibling ORDer)로 구성하고, XML 문서의 구조 정보를 표현하여 색인하는 방식도 있다.[3] 또한, 색인 파일의 명사 추출은 불용어 사전, 조사 사전, 조사결 어미사전, 어미사전, 조사/어미형 명사 사전 등을 이용한 명사를 추출한 후 발생 빈도 및 역문헌 빈도, 문서 길이에 따른 정규를 통해 [4,7,8] 용어 가중치를 부여하여 사용자가 검색시 가중치에 따른 방법도 있다.
따라서 본 연구에서는 [3]에서 제안한 문서구조 정보를 개선하여 XML 기반의 색인 파일을 설계하여 Local에서 검색을 보다 효율적으로 할 수 있도록 하고자 한다.

3. 구조 정보와 색인 구성

특정 엘리먼트에 대한 접근이 가능하고 엘리먼트 간의 관계를 구하기 위해 DTD의 논리적 구조 정보를 사용해서 XML 문서를 효율적으로 관리할 수 있는 구조 정보를 제안한다.

3.1. 구조 정보

XML 문서에 대한 DTD와 구조 정보는 다음과 같다.

(1) XML 문서 DTD

Local CD_ROM에 내장된 논문 정보에 대한 XML DTD 구조와 각 Element 타입의 ID는 <그림 1>와 같다.

```
<!DOCTYPE Doc [
<!ELEMENT Doc(head, Content)>
<!ELEMENT head(Title, Author*)>
<!ELEMENT Title (#PCDATA)>
<!ELEMENT Author (name, department)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT department (#PCDATA)>
<!ELEMENT Content (summary,
introduction, text, conclude)>
<!ELEMENT summary (#PCDATA)>
<!ELEMENT introduction (#PCDATA)>
<!ELEMENT text (#PCDATA)>
<!ELEMENT conclude (#PCDATA)> ]>
```

그림 1 XML DTD와 ETID

(2) 물리적 구조

XML 문서 대한 물리적 구조 정보는 다음 <표 1>과 같다.

Tag name	DID	ETID	SORD	SSORD	Type	Start Offset	End Offset	Content

표 1 XML 문서의 물리적 구조 정보

■Tag name : 엘리먼트 이름 또는 애트리뷰트 이름

■DID : 문서 고유 번호

■ETID : 엘리먼트들 간의 계층 정보를 표현하기 위해 EID(Element ID)를 부여하고 이것은 문서상의 엘리먼트들에 ETID(Element Type ID)를 부여함으로써 논리적인 구조를 나타내게 된다.

■SORD, SSORD : XML 문서에서는 동일한 엘리먼트들이 반복적으로 나타날 수 있는데 ETID만으로는 엘리먼트들의 반복적인 사용이 불가능해서 엘리먼트간의 현제 노드간 순서 정보를 나타내기 위한 SORD(Sibling ORDer)와 동일 타입의 엘리먼트들 간의 순서 정보를 나타내기 위해 SSORD(Same Sibling ORDer) 사용

■Type : 엘리먼트인지 애트리뷰트인지를 구별

■Start Offset, End Offset : 엘리먼트 및 애트리뷰트의 시작과 끝 위치

■Content : 해당 엘리먼트에 속하는 실제 내용이거나 애트리뷰트의 값

3.2. 구조 정보 및 색인 정보 추출

<그림 2>는 DTD에 대한 물리적 구조 정보를 추출하기 위한 과정이다.

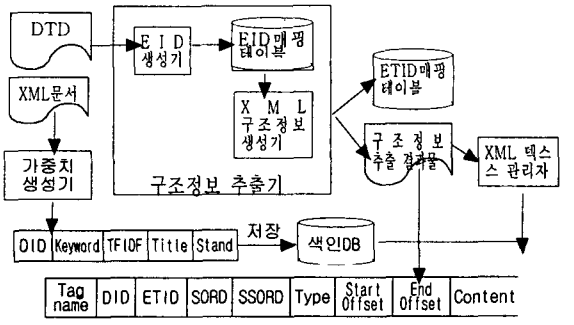

그림 1 구조 정보 및 색인 정보 추출기

여기서 EID 생성기는 DTD로부터 각각의 엘리먼트들에 대해 엘리먼트 타입을 추출하고 엘리먼트 이름과 EID를 사상시켜주는 매핑 테이블을 구성하고 XML 인덱스 관리자는 XML 문서에 대해 내용, 구조, 애트리뷰트, 혼합 검색을 지원하기 위한 색인 구조를 만들고 색인 정보를 관리한다.

3.3. 색인 정보와 색인 구성

(1) 단어 가중치 계산법

단어 가중치를 구성하는 요소는 단어 빈도(TF, Term Frequency), 역문헌 빈도 ((IDF, Inverse Document Frequency), 문헌 길이 정규화의 세 가지로 길이가 긴 문헌일수록 다른 문헌과의 유사도가 상대적으로 높아질 여지가 있으므로 문헌 길이를 반영한 정규화가 필요하다.

■단어 가중치 = $\frac{N}{SF_i}$ (N : 논문을 구성하는 전체 문장수, SF_i : 논문 내에서 용어 i가 출현한 문장의 수)

■표제어 가중치 =

$$\frac{\text{표제어가 포함된 문장이 요약 문장에 포함될 확률}}{\text{일반 문장이 요약 문장에 포함될 확률}}$$

■위치 가중치 =

$$\frac{\text{특정 위치의 문장에 요약문이 포함될 확률}}{\text{일반 문장이 요약 문장에 포함될 확률}}$$

(2) 물리적 구조 정보에 대한 인덱스는 <표 2>와 같다.

DID	Keyword	TFIDF	Title	Stand

표 2 물리적 구조 정보에 대한 인덱스 정보

■DID : 문서 고유 번호

■Keyword : 인덱스를 구성하게 될 용어

■TFIDF : Keyword의 출현 빈도

■Title : Keyword가 표제에 있을 때의 가중치

■Stand : Keyword가 특정 위치에 있을 때의 가중치

따라서 XML 구조+내용 검색이 이루어질 수 있도록 구조 검색은 관련 논문과 같이 ETID와 SORD, SSORD를 참조하여 특정한 엘리먼트에 대한 직접적인 접근이 가능하도록 하고 내용 검색은 질의와 매칭되는 색인 정보를 추출하여 해당 문서에서 keyword의 가중치를 고려하여 가중치에 의한 Sorting으로 이용자에게 제공한다.

4. 질의 처리

질의 처리 및 검색 결과는 내용 검색과 구조 검색을 구분하여 내용검색이 사용자로부터 요청될 경우는 XML 객체 관리자의 XML 인스턴스 관리기와 키워드의 가중치를 통해 저장매체에 저장되어 있을 문서 전체 혹은 일부분을 사용자에게 제공한다.

4.1. 질의 처리 절차

질의 처리 절차로는 먼저 질의어 입력를 입력하면 질의어에 대한 매핑 테이블에서 ETID와 색인 DB에 저장된 색인정보 DID, ETID, TFIDF, TITLE, Stand 정보를 추출한다. 추출된 정보중 EDIT를 참고하여 구조 정보를 추출하여 동일 타입의 엘리먼트들 간의 순서 정보를 나타낼 때는 SSORD 정보를 이용하여 새로운 구조 정보 추출하고 엘리먼트간의 형제 노드간 순서정보를 나타낼때는 SORD 정보를 이용하여 새로운 구조 정보를 추출한다. 문서에 포함된 질의어의 가중치를 비교하여 임의의 값 이상일때는 저장된 문서를 검색하여 사용자에게 제공한다.

다음 <그림 3>은 질의 처리 절차를 도식화했다.

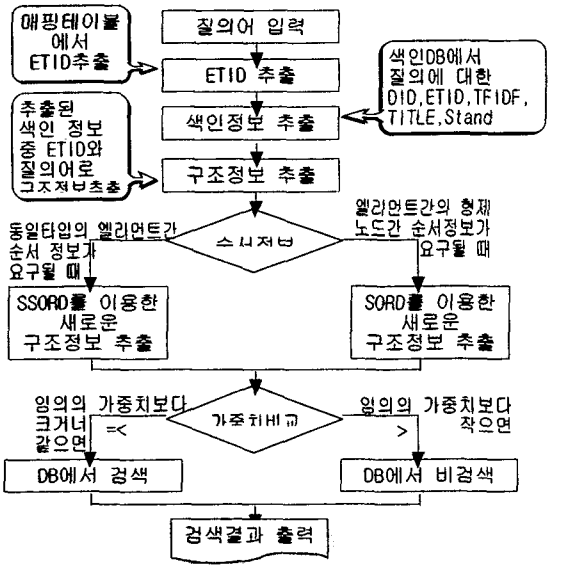

그림 2 질의 처리 절차

4.2. 질의 처리 결과

<그림 4>는 4.1에서 제안한 검색 알고리즘을 적용해서 정보처리학회 발표회의 논문 중 제목에 XML이 존재하는 논문을 검색한 결과이다.

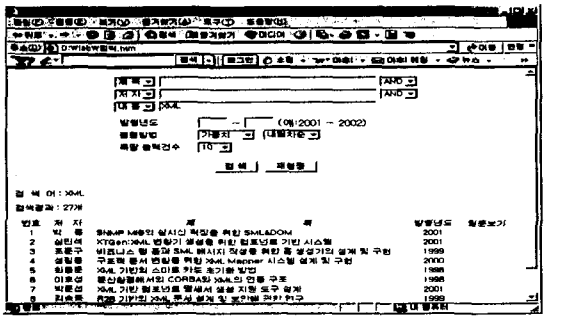

그림 3 검 색 결 과

5. 결론 및 향후 연구 과제

정보의 양이 증가함에 따라 적합한 문서를 검색하는데 우선순의를 결정하는데는 한계가 있지만 이 논문에서 제안한 색인 모델을 통해 사용자 질의에 대한 처리 과정을 통해 Local에서의 검색을 효율적으로 할 수가 있다. 기존의 XML 문서에 대한 구조 정보에 인덱스 정보를 추가하여 질의어에 대한 해당 문서의 가중치를 기준으로도 나타냄으로써 질의에 대한 연관성을 알 수 있으므로써 신뢰도를 향상할 수 있다. 향후 연구로서는 가중치를 부여한 내용+구조 색인을 실제 활용할 수 있도록 각 서브시스템을 구현 하는데 있다.

참 고 문 헌

[1] 신승호, 손충범, 유재수, "XML 문서의 효율적인 구조 검색을 위한 동적 색인 모델", 컴퓨터 정보 통신 연구, Vol.9 No.2 [2001] ,2001

[2] 김병진, 김두현, 홍도석, 김용성, " XML 기반한 Local 검색 시스템의 설계 및 구현", 한국 정보 과학회, 2002 추계 학술 발표 논문집, 2002

[3] 박종관, 강형일, 송충범, 유재수"XML 문서에 대한 효율적인 구조 기반 검색을 위한 색인 모델", 한국 정보 과학회, 2000 추계 학술 발표 논문집 pp.18-20, 2000

[4] 강승식,"지능형 정보검색을 위한 자동색인기법"

[5] Jon Bosak, "XML, Java, and the futere of the Web",http://sunsite.unc.edu/pub/sun-info/standa-rds/xml/why/xmlapps.htm/

[6] Richard Lander, "XML: The New Markup Wave",http://www.csclub.uwa.terloo.ca.u/relander/XML/Wave/xml_mw.html/

[7] Salton, G, and Buckley, C. " Term-weighting approaches in automatic text retrieval, information processing&Management, 24(5):513-523, 1988

[8] Singhal, A., et al."Pivoted document length normalization", Proceedings of the SIGIR, 96:21-29, 1996