링크 정보를 활용한 XML 문서의 검색*

문찬호, 강현철
중앙대학교 컴퓨터공학과
e-mail : {moonch,hckang}@rose.cse.cau.ac.kr

Retrieval of XML Documents Using Link Information

ChanHo Moon, Hyunchul Kang
Dept. of Computer Science and Engineering, Chung-Ang University

요 약

다양한 정보 형태를 가진 전자 문서의 효과적인 관리를 위해 XML 과 관련된 연구들이 활발히 진행되고 있다. 그러나 XML 과 관련된 대부분의 연구들은 XML 문서들을 저장, 관리 및 검색할 수 있는 XML 저장 관리 시스템을 대상으로 하고 있다. 본 논문에서는, 실제 Web 상에 분산되어 저장된 전자 문서들 중 하이퍼링크로 연결된 XML 문서들을 대상으로 사용자 질의에 대해 효율적인 검색을 지원해주기 위해, XML 링크 정보를 추출하여 참조하는 검색 질의 처리 방안을 제시한다. 이를 위해 링크 정보를 저장하는 링크 정보 관리 테이블의 구조, 링크 정보를 활용한 XML 문서의 검색 모델, XML 문서의 검색 질의 처리 방안, 그리고 링크 정보의 갱신을 질의 처리 중에 부가적으로 수행하는 방안을 기술하였다. 주어진 사용자 질의에 대한 처리 과정 중 링크와 관련된 질의 처리에 대해 추출된 링크 정보를 이용하여 부질의를 생성하고 처리할 수 있도록 하였다.

1. 서론

현재 Web 환경에서는 다양한 형태의 정보를 내포하고 있는 전자 문서들이 기하급수적으로 증가하고 있다. 이에 따라 전자 문서들에 대한 효과적인 관리 기능의 필요성이 대두되고 있다. 이러한 요구를 충족시키기 위해 XML (eXtensible Markup Language)이 등장하였다[1]. XML 은 인터넷을 구성하는 HTML 의 대안으로 W3C 에서 제안한 차세대 표준 마크업 언어이다. XML 의 특징을 살펴보면 다음과 같다.

- XML 은 간단하면서 다루기 쉽기 때문에 Web 상에서 제공하는 방대한 양의 전자 문서들을 XML 로 다양하게 표현할 수 있다. 특히 HTML 과는 달리 새로운 태그와 속성을 정의할 수 있다.

- 문서 구조의 검증이 필요한 응용개발 과정에서 XML 은 문법적 오류에 관한 판단을 문서 내에서 제공한다.

다양한 정보 형태를 가진 전자 문서의 효과적인 관리를 위해 XML 과 관련된 연구들이 현재 활발히 진행되고 있다. XML 문서들을 저장, 관리 및 검색할 수 있는 XML 저장 관리(repository) 시스템 개발에 대한 연구[2][3], XML 과 관련된 질의어에 관한 연구[4][5], 기존 데이타베이스 시스템에 저장된 데이타를 XML 로 변환하는 도구 개발에 관한 연구(관계 데이타베이스에서 XML 문서로, 객체 지향형 데이타베이스에서 XML 로 표현)[2][6] 등이 그 예이다. 이들 XML 과 관련된 연구들의 대부분은 XML 문서를 저장하고 있는 XML 저장 관리 시스템에 관한 것이다.

실제 Web 상에는 HTML 문서와 같이 XML 문서들도 서로 하이퍼링크로 연결되어 있는 상태로 분산되어 있다. 즉, 서로 다른 Web 사이트들 간에도 XML 문서에 대한 링크가 존재한다. 본 논문에서는, 실제 Web 상에 분산되어 저장된 전자 문서들 중 하이퍼링크로 연결된 XML 문서

* 본 논문은 정통부의 정보통신우수시범학교 지원 사업에 의한 것임.

들을 대상으로 사용자 질의에 대해 효율적인 검색을 지원해주기 위해, XML 링크 정보를 추출하여 참조하는 검색 질의 처리 방안을 제시한다. 주어진 사용자 질의에 대한 처리 과정 중 링크와 관련된 질의 처리에 대해 추출된 링크 정보를 이용하여 부질의를 생성하고 처리할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2 절에서는 본 논문과 관련된 기존 연구를 살펴보고, 3 절에서는 링크 정보를 활용한 XML 문서의 검색 질의 처리 방안을 제시하고, 4 절에서 결론을 맺는다.

2. 관련 연구

XML 은 기존의 정보와는 달리 논리적 구조를 가지고 있다. 이러한 XML 문서를 대상으로 구조 정보를 이용하여 문서를 효율적으로 관리하기 위한 XML 저장 관리 시스템 개발에 관한 연구들은 XML 문서, DTD(Document Type Definition)의 구조 정보 및 다양한 미디어 저장, 관리에 대한 처리를 대상으로 하고 있다. 또한 XQL(XML Query Language)[4]이나 XQL-QL[5]은 XML 문서의 구조적인 특성을 반영한 구조/내용/속성 기반 검색들을 지원하고 있다. 그리고 XML 문서와 관련된 질의 처리를 효과적으로 수행하기 위한 색인 기법에 관한 연구[7]도 진행중이다. 그러나 Web 을 대상으로 각 사이트에 분산 저장되어 있는 XML 문서들을 대상으로 한 질의 처리 기능에 대해서는 연구가 미비하다.

Web 상에 분산된 정보에 대한 질의 처리를 위한 연구도 진행 중에 있다[8][9]. 특히, Web 환경에서 이질적인 정보에 대해 데이타의 추출 및 통합 기능을 위한 wrapper-mediator 기반 관련 연구[10]가 활발히 진행되고 있다. wrapper-mediator 기반의 연구는 Web 사이트, 화일 시스템, 데이타베이스 시스템에 산재된 데이타 소스를 대상으로 질의 처리를 수행하고자 하는 것이다.

3. 링크 정보를 활용한 XML 문서 검색

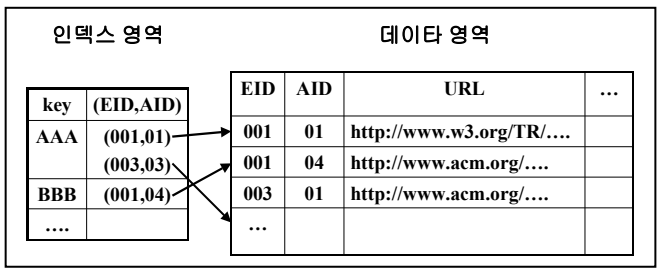
3.1 개요

Web 상에 분포되어 저장된 전자 문서들은 링크를 통해 서로 간의 참조를 허용하고 있다. 본 논문에서 제안하는 링크 정보를 활용한 XML 문서 검색은, 이러한 전자 문서들 중 XML 문서들을 대상으로 효율적인 검색을 지원해주기 위해 링크를 추출, 참조하는 질의 처리 기능을 제공함을 목적으로 한다. 링크 정보를 활용한 XML 문서 검색은 XML 문서들 간의 링크 정보를 활용하여 같은 사이트에 포함된 XML 문서 뿐만이 아니라 서로 다른 여러 사이트에 포함된 XML 문서들에 대한 검색 기능도 지원한다. 링크 정보를 활용한 XML 문서 검색을 지원하기 위해 본 절에서는, 링크 정보를 저장하는 링크 정보 관리 테이블의 구조, 링크 정보를 활용한 XML 문서의 검색 모델, XML 문서의 검색 질의 처리 방안, 그리고 링크 정보의 갱신을 질의 처리 중에 부가적으로 수행하는 방안을 제시한다.

3.2 링크 정보 관리 테이블

현재 대부분의 Web 사이트 관리자들은 자신들이 관리하는 사이트 내에 많은 전자 문서(특히, HTML 문서)를 저장, 관리하고 있다. XML 문서도 HTML 문서와 마찬가지로 한 사이트에 많이 포함되어 있다. 한 사이트 전체를 대상으로 한 검색 질의가 발생할 경우 사이트 내 전체 XML 문서를 검색하여 처리해야 한다. 이러한 처리는 검색 시간이 오래 걸린다는 단점이 있기 때문에 대부분의 XML 문서를 대상으로 하는 검색 시스템들은 색인 정보를 활용하여 검색 질의 처리를 수행한다. 링크 정보를 활용한 XML 문서 검색 질의의 경우에도 이와 같은 색인 정보를 이용한다. (그림 1)은 링크 정보를 활용한 질의 처리에 사용되는 링크 정보 관리 테이블을 나타낸 것이다.

링크 정보 관리 테이블은, 참조되는 XML 문서의 URL 을 저장하는 데이타 영역과 데이타 영역의 검색을 효율적으로 수행하기 위한 인덱스 영역으로 크게 나뉜다. 링크 정보 관리 테이블은, XML 문서를 관리하는 한 사이트에 포함된 모든 XML 문서들을 대상으로 XML 문서 내 링크 정보(URL)를 가진 엘리먼트(element)와 애트리뷰트(attribute), URL, 그리고 URL 을 가진 키(검색 시 사용되는 키워드)값들로 구성된다. (그림 1)에 나타낸 것과 같이 XML 문서 내 링크 정보는 같은 사이트에 포함된 XML 문서 뿐만이 아니라 다른 사이트에 저장된 XML 문서를 포함하고 있다.



(그림 1) 링크 정보 관리 테이블

XML 문서 내 참조되는 URL 의 위치는 URL 이 포함된 엘리먼트 ID(EID: element identifier)와 애트리뷰트 ID(AID: attribute identifier) 쌍으로 나타낸다. 이는 XML 문서를 대상으로 하는 질의 처리 중 사용되는 색인 기법[7]을 활용한 것이다. EID 와 AID 쌍은 XML 문서 내에서 유일하게 할당된다. 링크 정보 관리 테이블의 데이타 영역은 (EID, AID, URL,…)의 엔트리들로 구성된다. 인덱스 영역은 URL 과 연관된 키값과 키값이 저장된 XML 문서

의 EID 와 AID, 그리고 URL 를 포함하고 있는 링크 정보 관리 테이블의 데이타 영역 엔트리를 가리키는 포인터로 구성된다. (그림 1)은 키값이 AAA 인 키워드에 대해서는 2 개의 URL 참조 링크가 있으며 그 중 하나의 URL 은 http://www.w3.org/TR/…임을 나타내고, BBB 인 키워드에 대해서는 1 개의 URL 참조 링크가 있으며 그 URL 은 http://www.acm. org/…임을 나타낸 것이다.

XML 문서는 빈번히 추가되고 삭제될 수 있다. 또한 XML 문서의 구조 또한 동적으로 변화하기 때문에 엘리먼트와 애트리뷰트의 변화가 자주 발생한다. 이와 관련해서 링크 정보 관리 테이블도 적절히 수정되어야 한다. 링크 정보 관리 테이블의 링크 정보 수정 시점에 대해서는, 상기 변화 발생 시 즉각 수정을 반영하는 방안과 지연한 뒤 적절한 시점에 수정을 반영하는 방안이 있을 수 있다. 그러나 XML 문서의 변화는 빈번히 발생하므로 전자의 방안은 비효율적이다. 그러한 이유로 본 논문에서는 링크 정보 관리 테이블의 링크 정보 수정은 질의 처리 시 수행하는 방안을 채택한다. 즉, 질의 처리 중 관련된 XML 문서의 링크가 새로이 추가된 경우에는 링크 정보 관리 테이블에 해당 링크 정보를 추가하고, XML 문서의 링크가 삭제된 경우에는 링크 정보 관리 테이블에 해당 링크 정보를 삭제한다

3.3 링크 정보를 활용한 XML 문서의 검색 모델

(그림 2)는 Web 상에 XML 문서를 저장하고 있는 각각의 사이트를 대상으로 링크 정보를 활용한 XML 문서 검색 모델을 나타낸 것이다. 각 사이트는 XML 문서들을 저장하고 있는 XML 소스(source)와 사용자 질의를 처리할 수 있는 검색 모듈로 구성된다. 질의 처리 중 링크 정보를 활용한 XML 문서 검색 모듈의 구성은 다음과 같다.

- DTD 관리 모듈 : XML 소스에 저장된 XML 문서들의 DTD 를 자동으로 추출하여 DTD 저장소(repository)에 저장한다.

- 링크 정보 관리 모듈 : XML 소스에 저장된 XML 문서들 중 링크 정보를 갖는 내용을 추출하여 링크 정보 관리 테이블에 저장한다.

- 질의 처리 모듈(Query Processor) : 사용자 인터페이스를 통해 입력 받은 질의를 파싱하여 해당 XML 소스에 저장된 XML 문서를 대상으로 질의 처리를 수행한다. 질의 처리 수행 중 XML 의 논리적 구조를 필요로 하는 경우 DTD 저장소에 저장된 DTD 를 참조하고, 링크 정보를 필요로 하는 경우 링크 정보 관리 테이블을 참조한다. 다른 XML 소스에 대한 검색이 필요한 경우 부질의(사용자 인터페이스로 입력받은 질의 중 링크 정보를 활용한 질의만을 처리하도록 수정한 질의) 생성 모듈에 부질의 생성을 요청한다. 이 경우 여러 XML 소스에 대한 검색이 필요한 경우에는 각 XML 소스를 대상으로 부질의 생성을 요청한다.

- 부질의 생성 모듈(Subquery Generator) : 다른 XML 소스에 저장된 XML 문서를 대상으로 부질의를 생성한다.

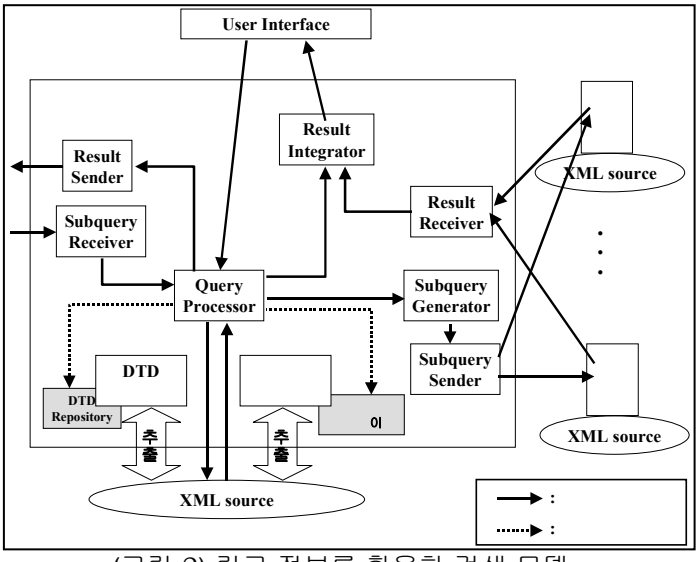
- 부질의 전달 모듈(Subquery Sender) : 부질의 생성 모듈에 의해 생성된 부질의들을 해당 XML 소스의 질의 처리 모듈로 부질의를 전달한다.

- 부질의 수신 모듈(Subquery Receiver) : 타 검색 모듈로부터 전달된 부질의를 수신하여 자신의 검색 모듈 내 질의 처리 모듈로 전달한다.

- 결과 전달 모듈(Result Sender) : 타 검색 모듈로부터 전달된 부질의의 처리된 결과를 해당 검색 모듈로 되돌려 준다.

- 결과 수신 모듈(Result Receiver) : 타 검색 모듈에 요청한 부질의의 결과를 수신하여 결과 합성 모듈로 전달한다.

- 결과 합성 모듈(Result Integrator) : 해당 사이트에 대한 XML 검색 결과와 타 사이트에 대한 XML 검색 결과를 합성하여 사용자에게 전달한다.



(그림 2) 링크 정보를 활용한 검색 모델

3.4 링크 정보를 활용한 XML 문서의 검색 질의 처리

사용자 인터페이스를 통해 검색 질의를 받은 경우, 질의 처리 모듈은 질의 파싱을 통해 사용자 질의로부터 검색하고자 하는 키값을 추출한다. 다음은 검색 질의의 예를 나타낸 것이다.

A 사이트 내의 문서 저자 중 “A. Einstein”인 사람의 홈페이지를 찾아 “A. Einstein”이 쓴 논문들을 찾으시오.

이와 같은 질의는 Web 을 대상으로 빈번히 발생할 수 있다. 이 질의를 3.3 절에서 제시한 XML 문서 검색 모듈에 의해 처리될 수 있는 질의어(기존 데이타베이스 시스템에서 찾아볼 수 있는 구조 기반의 질의에 Web 검색 엔진의 내용 기반 검색을 지원하는 Web 관련 질의어인 W3QL[11]의 문법을 확장한 형태로 SQL 형태를 가짐)로 나타내면 다음과 같다.

```
select Y.document.biblio/*
  from X in A
        where
             Y in X.document.name[@href]
        and
             X.document.name = “A. Einstein”
```

XML 문서를 대상으로 엘리먼트 계층 간의 구분은 점(.)으로 표현하였고, XML 문서의 애트리뷰트는 대괄호([])로 표현하였다. 그리고 XML 문서의 하이퍼링크 구조는 @href 로 표현하였다. 그리고 질의어 첫줄의 /*는 모든 엘리먼트의 내용을 검색하라는 의미를 갖는다. 위의 질의에서 링크와 관련된 키값은 “A. Einstein”이므로 질의 처리 모듈은 질의 처리 과정에서 먼저 “A. Einstein”을 추출한다.

질의 처리 모듈은 추출된 키값으로 링크 정보 관리 테이블의 인덱스 영역 엔트리를 찾는다. 찾아진 엔트리를 통해 키값에 대한 URL 정보를 얻게 된다. 이때, 만약 사용자가 요구하는 키값에 대한 엔트리가 링크 정보 관리 테이블의 인덱스 영역에 없는 경우는 사용자 질의를 요청받은 해당 XML 소스 내 키값에 대한 URL 정보가 없음을 의미한다. 이는 링크 정보 관리 테이블을 통해 XML 소스 내 전체 XML 문서를 검색할 필요가 없음을 의미한다.

찾아진 URL 정보를 얻은 후, 질의 처리 모듈은 부질의 모듈에게 부질의 생성을 요청한다. 앞에서 예시한 질의에 대한 부질의 탬플릿은 다음과 같다.

```
select Y.document.biblio/*
from Y in B
```

B 는 앞에서 기술한 키값(“A. Einstein”)을 통해 찾아낸 URL 을 나타낸다. 생성된 부질의는 B URL 에 해당하는 XML 문서의 biblio 엘리먼트 검색을 수행하라는 의미를 나타낸다. “A. Einstein”과 관련된 URL 이 많이 있는 경우에는 해당 URL 당 부질의를 생성하게 된다. 부질의는, 키값(“A. Einstein”)에 대한 URL 에 해당하는 XML 문서가 속한 XML 소스(B 사이트)의 질의 처리 모듈로 전달되어 XML 문서를 대상으로 질의 처리된다. 부질의의 처리된 결과는 결과 합성 모듈로 전달되고, 결과 합성 모듈은 여러 XML 소스에서 전달된 결과를 합성하여 사용자 인터페이스를 통해 사용자에게 전달한다.

3.5 질의 처리 중 링크 정보 테이블 갱신

3.2 절에서 기술한 것과 같이 링크 정보의 갱신은 질의 처리 중에 부가적으로 수행한다.

• 링크 정보의 추가

XML 소스에 새로운 XML 문서가 추가되거나 기존의 XML 문서에 대해 동적으로 엘리먼트나 애트리뷰트가 추가되어 갱신된 경우, 추출 과정을 통해 검색에서 사용될 키값, URL 과 키값이 속하는 엘리먼트, 애트리뷰트를 얻어 링크 정보 관리 테이블에 저장한다.

• 링크 정보의 삭제

XML 소스에서 기존의 XML 문서가 삭제되거나 엘리먼트나 애트리뷰트가 삭제된 경우, 추출 과정을 통해 삭제되는 URL 과 키값을 포함하는 엘리먼트와 애트리뷰트를 얻어 링크 정보 관리 테이블 내 해당 엔트리를 삭제한다.

4. 결론

본 논문에서는, 실제 Web 상에 분산되어 저장된 전자문서들 중 하이퍼링크로 연결된 XML 문서들을 대상으로 사용자 질의에 대해 효율적인 검색을 지원해주기 위해, XML 링크 정보를 추출하여 참조하는 검색 질의 처리 방안을 제시하였다. 주어진 사용자 질의에 대한 처리 과정 중 링크와 관련된 질의 처리에 대해 추출된 링크 정보를 이용하여 부질의를 생성하고 처리한다. 링크 정보를 활용한 XML 문서 검색을 지원하기 위해 본 논문에서는, 링크 정보를 저장하는 링크 정보 관리 테이블의 구조, 링크 정보를 활용한 XML 문서의 검색 모델, XML 문서의 검색 질의 처리 방안, 그리고 링크 정보의 갱신을 질의 처리 중에 부가적으로 수행하는 방안을 기술하였다.

참고문헌

[1] T. Bray et al., “Extensible Markup Language (XML) 1.0,” http://www.w3.org/TR/1998/REC-xml-19980210, 1998.

[2] 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.

[3] L. Liu et al., “Continual Queries for Internet Scale Event-Driven Information Delivery,” IEEE Trans. on Knowledge and Data Engineering, Vol. 11, No. 4, 1999, pp. 610-628.

[4] J. Robie et al., “XML Query Language (XQL),” http://www.w3.org/TandS/QL/QL98/pp/xql.html, 1998.

[5] A. Deutsch et al., “XML-QL: A Query Language for XML,” http://www.w3/TR/NOTE-xml/, 1998.

[6] “Transforming Relational Databases into XML Documents”, http://www.informatik.fh-wiesbaden.de/~turau/ DB2XML/index.html.

[7] 이계준 외 2 명, “XML 문서의 검색을 위한 효율적인 색인 기법과 질의 언어(TQL)의 설계,” 한국정보과학회, ’99 년 가을 학술발표논문집, 26 권, 2 호, 1999, pp. 57-59.

[8] M. Fernandez et al., “Catching the Boat with Strudel: Experience with a Web-Site Management System,” SIGMOD Record, Vol. 27, No. 3, 1998, pp. 414-425.

[9] A. Sahuguet and F. Azavant, “W4F: a WysiWyg Web Wrapper Factory,” http://cheops.cis.upenn.edu/~sahuguet /WAPI/wapi.ps.gz, 1999.

[10] D. Florescu et al., “Database Techniques for the World-Wide Web: A Survey,” SIGMOD Record, Vol. 27, No. 3, 1998, pp. 59-74.

[11] D. Konopnicki and O. Shmueli, “W3QS: A Query System for the World-Wide Web,” Proc. Int’l Conf. on VLDB, 1995, pp. 54-65.