XLink를 이용한 XML 문서의 인덱싱 기법

김선경°, 김성완[1], 정헌석, 이재호[2], 임해철
홍익대학교 컴퓨터공학과, [1]삼육의명대학 컴퓨터정보과, [2]인천교육대학교 컴퓨터교육과
{skkim°, lim}@cs.hongik.ac.kr, swkim@syu.ac.kr, jhlee@mail.inue.ac.kr

Indexing Scheme for XLink in XML Documents

Sun Kyung Kim°, Sung Wan Kim[1], Hun-Suk Chung, Jaeho Lee[2], HaeChull Lim
Dept. of Computer Eng., Hong Ik Univ., [1]Dept. of Computer and Information, Sahm Yook College.,
[2]Dept. of Computer Education, Inchon Nat'l Univ. of Education

요 약

인터넷의 급속한 발전과 더불어 대량의 정보를 효과적으로 표현 및 교환할 수 있는 표준으로 XML이 제정된 이후, XML 문서의 저장과 검색을 위한 많은 연구들이 진행되고 있다. 한편, XML 문서간의 관계를 표현하기 위한 XLink가 제정되면서, XLink로 표현된 링크 정보를 이용하여 문서들을 효과적으로 검색할 수 있는 정보 검색 시스템에 대한 연구가 진행되고 있지만 그 성과가 미흡하다. 따라서 첫째, 본 논문에서는 링크 정보를 가지고 있는 XML 문서의 데이터 모델을 정의하고, 문서간 링크 정보가 가져야 할 링크 참조 무결성을 제안하였다. 둘째, 링크 정보를 이용한 질의 처리를 위해 제안한 모델과 최신 XLink 표준을 준수하여 테이블 형식의 링크 정보 인덱스 구조를 설계하였다.

1

인터넷의 발전과 www의 보편화로 인해 교류되는 정보의 양이 크게 증가함에 따라, 대량의 정보를 보다 효과적으로 저장 및 관리할 수 있는 새로운 데이터 표준으로 XML (eXtensible Markup Language)이 제안되었다[1]. 이에 따라, XML 문서에 대한 저장과 인덱싱에 대한 연구가 특히 활발하게 진행되고 있다[2,3]. 그러나, 이러한 연구들은 독립적인 XML 문서를 단일 대상으로 하는 경우가 대부분이기 때문에 XML 문서간의 관계를 정의한 XLink의 개념을 지원하지 못하는 한계가 있다. 이러한 이유로 여러 문서간의 관계를 이용하여 사용자의 질의를 보다 효과적으로 검색하기 위하여 XLink를 이용한 인덱싱 기법에 대한 연구가 필요하다. 현재, 초기 단계로 연구 결과가 미약하다.

기존 정보 검색 시스템의 인덱싱 기법들[2][3]은 주로 단일 문서만을 고려하였기 때문에 링크 된 문서를 검색의 대상으로 고려하지 않은 인덱스였다. 또한, 기존의 인덱싱 연구는 최신의 XLink의 표준[4]의 특징을 모두 지원하지 않는다는 단점이 있다. 따라서, 본 논문에서는 XLink의 애드리뷰트 중 링크의 의미정보를 포함하고 있는 title과 inbound link를 지원하기 위한 인덱스를 설계하였다.

2. 관련 연구

2.1. 가중치를 이용한 문서 인덱싱 기법

[5]에서는 XLink의 링크 속성 중 actuate와 show를 이용한 가중치 부여 방법을 사용한다. 이를 이용해 링크 된 문서의 중요성을 판단해서 우선 순위를 부여한다.

인덱싱 시에는 역화일을 이용한 색인어 기반의 포스팅(posting)을 생성하고, 문서의 링크 정보는 <그림 1>과 같은 링크 정보 화일에 따로 저장해 놓는다.

element	id	frequency	href	Remote_link	Insert_link

<그림 1> 링크 정보 화일

문서에서 인덱싱이 끝나면, 임시 화일에 들어있는 링크 정보를 이용하여 링크 된 문서에서의 엘리먼드 정보를 포스팅에 삽입한다. 이러한 방법으로 링크 된 모든 문서들을 인덱싱 하는데, 링크 식별자 테이블과 원격 문서 안의 링크에 대한 가중치 부여 테이블에서 가중치 값이 '0'일 때까지 이 과정이 반복적으로 이루어진다.

하지만, Remote_Link 값을 결정해 주는 판단 기준이 모호하며 수동적이다. 또한, 가중치를 부여할 때도 단순링크(simple link)와 확장링크(extended link)에서의 인라인 링크(inline link)만을 가정하였기 때문에 최신의 XLink 표준을 따르는 문서의 검색에 있어서 문제가 된다. 그리고, 문서 혹은 엘리먼드 단위의 검색이 용이하지 않다.

2.2. XQL(XML document Query Language)

[6]에서는 기존 XML 문서의 데이터 모델에 링크 정보를 포함하고 있는 XML 문서의 데이터 모델을 추가하여 XLink로 표현된 XML 문서의 데이터 모델을 제안하였다. 그러나, 제안된 데이터 모델은 HTML에서처럼 단방향의 링크만을 지원하였고, XLink가 가지는 확장된 형태의 링크를 표현하지 못하는 단점이 있다.

3. XLink를 이용한 인덱싱 기법

기존 XLink의 애드리뷰드를 이용한 인덱스는 Actuate, Show등의 애트리뷰트만을 이용하였기 때문에 실제 의미정보를 가지고 있는 애드리뷰드에 대한 인덱싱은 이루어지지 않았다. 또한 링크 되어 있는 여러 문서 중에서 어떠한 문서를 색인 할 지에 대한 명확한 판단 기준이 분명하게 정의되지 않았다[5]. 한편, [2]에서는 한 문서에 대한 인덱싱은 가능하나 웹 환경에서와 같은 여러 문서의 인덱싱은 이루어지지 않았다[7]. 따라서 본 절에서는 타입과 애트리뷰트에 대한 인덱싱과 outbound Link와 inbound link 모두를 지원하는 인덱싱 기

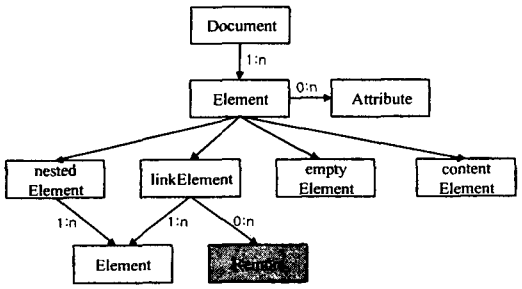
본 연구는 정보통신연구진흥원 대학기초연구지원사업(과제번호 : C1-20011-122-3)의 지원을 받았음

법을 제시한다.

3.1. 데이터 모델

3.1.1 구조

XLink로 표현된 링크 정보를 고려한 XML 문서의 데이터 모델을 위한 구조적 측면은 <그림 2>와 같다.

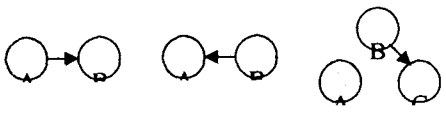


<그림 2> 링크 정보를 포함한 XML 문서의 데이터 모델

하나의 문서(Document)는 문서의 논리적 구조 정보를 표현하는 엘리먼드(Element)와 1:n의 관계를 갖는다. 엘리먼드는 애드리뷰드(Attribute)와 0:n의 관계에 있고, 중첩 엘리먼트(nestedElement), 빈 엘리먼드(emptyElement), 내용 엘리먼드(contentElement)와 XLink로 표현된 링크 정보를 가질 수 있다. <그림 2>에서 엘리먼드와 링크 관계에 있는 대상을 Remote로 표현하였으며, 이는 원격 자원(remote resource)를 의미한다. 여기서 자원이란 문서, 이미지, 오디오, 화일 등의 도달 가능한 정보의 단위를 의미[4]하며, 본 논문에서는 XML 문서 또는 엘리먼드로 그 의미를 한정하겠다. 링크 정보를 포함하고 있는 엘리먼드는 원격 자원과 0:n의 관계를 가진다. 내용은 엘리먼드가 포함하고 있는 텍스트 값이고, 애트리뷰트는 엘리먼트의 속성이다. 그리고, 중첩 엘리먼드는 엘리먼드 안에 내포되어 있는 엘리먼드이다.

3.1.2 링크 참조 무결성

링크는 도달 불가능한 자원(resource)과 관계성을 가질 수 없다. 도달 불가능한 자원은 링크 정보에 명시된 자원의 위치에 해당 자원이 없는 것을 말한다. 본 논문에서는 XLink[4]에 근거하여 3가지 링크 참조 무결성(link referential integrity)을 정의한다.



a. outbound link b. inbound link c. third-party link
<그림 3> 링크 참조 관계 예

A. outbound 링크 참조 무결성

지역 자원에서 원격 자원으로의 링크 참조 관계에서 지역 자원의 링크 정보는 NOT NULL이고 유효한 원격 자원의 위치 정보를 가져야한다. <그림 3>의 a와 같이 A→B로 링크 된 경우, B에 대한 위치 정보를 가지고 있는 A의 링크 정보 값은 널 값을 가질 수 없고, 반드시 B에 대한 실제 위치 주소 값을 가져야 한다. 여기서 실선으로 표시된 타원은 지역 자원을 의미하며, 점선으로 표시된 타원은 원격 자원을 의미한다.

B. inbound 링크 참조 무결성

원격 자원에서 지역 자원으로의 링크 참조 관계에서 지역 자원의 링크 정보는 NOT NULL이고 유효한 원격 자원의 위치 정보를 가져야한다. <그림 3>의 b와 같이 B→A로 링크 되어 있을 경우, A에 정의된 위치 정보 값은 B에 대한 유효한 위치 정보 값을 가져야 한다. 즉, B는 A의 위치 정보 값에서 명시한 위치에 반드시 존재해야 한다.

C. third-Party 링크 참조 무결성

원격 자원에서 원격 자원으로의 링크 참조 관계를 정의하는 지역 자원의 링크 정보는 NOT NULL이거나 유효한 원격 자원의 위치 정보를 가져야한다. <그림 3>의 c와 같이 B→C로 링크 되어 있고, A가 이에 대한 링크 정보를 가지고 있을 경우, A에 정의된 링크 정보 값은 반드시 유효한 B와 C의 위치 정보 값을 가져야 한다. 즉, B, C는 A의 링크 정보 값에서 명시한 위치에 반드시 존재해야 한다.

3.1.3 링크 정보 화일

본 절에서는 XML 문서에서 링크 정보를 고려하여 여러 문서간의 관계성을 이용한 검색 처리를 위한 인덱스 구조에 대해 설명한다. 인덱스의 기본적인 구조는 <그림 4>와 같다.

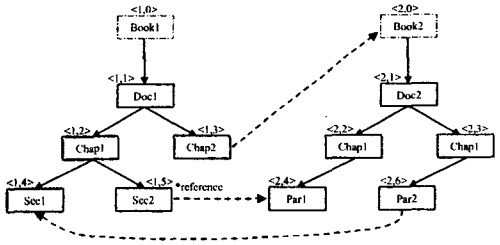
fromName	toName	fromID	toID	title

<그림 4> 링크 정보 화일의 구조

문서의 엘리먼드는 <DID, UID>쌍을 엘리먼드의 고유 식별자로 이용한다. DID는 문서 식별자이며, UID는 한 문서내의 엘리먼트에 할당되는 정수의 식별자로 모두 숫자 값을 갖는다.

<그림 4>에서 fromID는 링크 참조 관계가 시작되는 자원에 대한 식별자를 의미하며, toID는 링크 참조 관계의 목표가 되는 자원에 대한 식별자를 의미한다. fromID와 toID는 <DID, UID> 쌍을 값으로 갖는다. 여기서 자원이 문서 단위인 경우 UID 값은 '0' 값을 갖는다.

fromName은 fromID가 가리키고 있는 자원의 이름이고, toName은 toID가 가리키고 있는 자원의 이름이다. 단, 문서 단위의 링크 참조 관계의 경우 fromName와 toName은 문서의 화일 이름을 값으로 가진다. title은 fromName의 값인 엘리먼트가 가지는 애트리뷰드 중 링크를 사람이 식별할 수 있는 의미 정보를 가진 title 애드리뷰트의 내용을 의미한다. <그림 5>는 링크 참조 관계가 존재하는 두 문서간의 상태를 드리 형태로 나타낸 것이다. 여기서 방향을 갖는 점선으로 표시된 간선은 링크 관계를 나타낸다. *는 엘리먼드의 애트리뷰트 중 title 값을 의미한다.



<그림 5> 링크 관계가 포함된 두 문서

본 논문에서는 모든 문서가 가상의 루드 노드를 가지고 있다고 가정한다. 가상의 루트 노드에 대한 식별자는 <DID, UID>를 갖는

데, DID는 문서 번호이고, UID는 '0' 값을 갖고, 노드의 이름은 화일 이름과 동일한 이름을 갖는다. 이는 문서 단위의 검색에 이용한다.

본 논문에서 제안한 인덱싱 과정은 XML 문서를 파싱하여 각 엘리먼트에 식별자를 할당한 후, 링크 정보가 정의된 엘리먼트 혹은 문서를 대상으로 <그림 4>에 각 필드에 대한 정보를 추출하여 인덱스 테이블을 완성한다. <표 1>은 <그림 4>에 대해 생성된 인덱스 테이블이다.

fromName	toName	fromID	toID	title
Chap2	Book2	<1,3>	<2,0>	
Sec2	Par1	<1,5>	<2,4>	reference
Par2	Sec1	<2,6>	<1,4>	
...	...	...	...	...

<표 1> 제안한 링크 정보 화일의 예

<표 2>는 [5]에서 제안된 인덱싱 과정을 <그림 1>에 적용하여 생성한 링크 정보 화일의 예이다.

element	id	frequency	href	Remote_link	Insert_link
chap2	1	α	Book2.xml	Y	
Sec2	2	α	Book2.xml	Y	
Par2	2	α	Book1.xml/Doc1 /Chap1/Sec1	N	6
...	...	...	...	...	...

<표 2> 기존 링크 정보 화일의 예

3.2. 질의 예제

본 절에서는 본 제안한 인덱스와 [5]에서 제안한 인덱스를 이용한 몇 가지 대표적인 링크 질의 예와 이에 관한 처리 과정을 비교 설명한다.

A. 특정 엘리먼드가 링크하고 있는 원격 자원 검색
예) Chap2가 링크하고 있는 문서를 검색하라.
① 기존 링크 정보 화일
'Chap2'가 링크하고 있는 문서를 찾기 위해서는 href 정보를 이용하여야 한다. 여기서는 'Book2.xml'이 검색된 결과로 반환된다. 하지만, href 정보는 링크하고 있는 리소스의 위치 및 경로 정보로서 링크의 대상이 되는 문서 혹은 엘리먼트 단위에 대한 보다 세부적인 검색이 용이하지 않다.
② 제안된 링크 정보 화일
링크 정보 화일에서 fromName의 값이 'Chap2'이고, toID의 UID 값이 '0'인 엔트리들을 추출한다. 추출된 엔트리들에 대해 toName 필드의 값만 추출하면 최종 결과 문서들의 집합이 된다. 여기서는 Chap2가 링크하고 있는 문서는 'Book2'인 것을 알 수 있다.

B. 특정 엘리먼트를 링크하고 있는 원격 자원 검색
예) Sec1을 링크하고 있는 문서의 엘리먼드를 검색하라.
① 기존 링크 정보 화일
링크 정보 화일에서 'Sec1'을 링크하고 있는 문서의 특정 엘리먼트를 검색하기 위해서는 href 정보를 보고 판단해야 한다. 하지만, 앞의 예제에서 언급한 것처럼 href 정보만으로는 엘리먼드 단위의 검색이 어렵다. 따라서, 'Sec1'을 링크하고 있는 모든 문서의 엘리먼드를 효과적으로 검색할 수 없다.
② 제안된 링크 정보 화일
링크 정보 화일에서 toName의 값이 'Sec1'인 엔드리들을 추출한다. 추출된 엔트리들의 fromName과 fromID의 값을 최종 결과 집합으로 반환한다. 여기서는 Sec1을 링크하고 있는 문서는 식별자가 <2, 6>인 Par2 엘리먼트인 것을 알 수 있다.

C. title 정보를 이용한 검색
예) 'reference' 타이틀을 가진 링크 설정된 두 개의 엘리먼트를 검색하라.
① 기존 링크 정보 화일
링크 관계를 설명하는 title 애드리뷰드 정보를 저장하고 있지 않기 때문에 검색이 불가능하다.
② 제안된 링크 정보 화일
링크 정보 테이블의 title에서 값이 'reference'인 엔드리들을 추출한다. 추출된 엔트리들에 대해 fromName, fromID과 toName, toID 필드의 값을 최종 결과 집합으로 반환한다. 여기서는 식별자가 <1,5>인 'Sec2'인 엘리먼드와 식별자가 <2,4>인 'Par1' 엘리먼드가 검색된다.

4. 결론 및 향후 연구

본 논문에서는 링크 정보를 가지고 있는 XML 문서의 데이터 모델을 정의하고, 문서간 링크 정보가 가져야 할 참조 무결성을 제안하였다. 또한, 기존의 링크 정보를 이용한 인덱싱 기법이 가지는 최신 XLink의 표준의 특징을 모두 지원하지 않는 문제점을 해결하도록 링크 정보만을 따로 저장하고 있는 링크 정보 테이블 제안하였다. 또한, 대표적인 몇 가지 예제를 통해 링크 정보를 포함한 질의 처리 과정을 보였다.

향후 연구로는 보다 풍부한 링크정보를 이용한 인덱스의 확장과 보다 효율적인 검색이 가능하도록 Name 정보를 이용한 인덱싱 기법의 개발이 요구된다.

<참고문헌>

[1] T. Bray et al., "Extensible Markup Language (XML) 1.0 (Second Edition)", http://www.w3.org/TR/REC-xml, 2000

[2] Dongwook Shin, et al., "BUS: An Effective Indexing and Retrieval Scheme in Structured Documents", Proc.of the 3rd ACM Int'l Conference on Digital Libraries, 1998

[3] Q. Li, B. Moon, "Indexing and Querying XML Data for Regular Path Expressions", VLDB 2001

[4] S. DeRose et al, "XML Linking Language (XLink) Version 1.0", http://www.w3.org/TR/xlink/, 2001

[5] 김은정, 배종민, "XML 링크정보를 이용한 정보 검색 색인 기법의 설계", 한국정보처리학회 논문지, 제7권 제7호, 2000.7

[6] 김용훈, 이강찬, 이규철, "링크 검색을 지원하는 XML 문서 질의 언어(XQL)의 설계", 한국정보과학회 가을 학술발표논문집, Vol.25.No.2, 1998

[7] Timothy Arnold-Moore, Ron Sacks-Davis, "Models for Structured Document Database Systems", Markup Technologies, 1998