제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

XML 링크의 의미 속성을 이용한 개선된 검색 시스템을 위한 색인 기법에 대한 연구

김은정
부산외국어대학교 컴퓨터전자공학부
ejkim@taejo.pufs.ac.kr

A Study on Indexing Method for Advanced Retrieval System using Semantics Attributes in XML Links

Eun-Jung Kim
Dept. of Computer Engineering, Pusan University of Foreign Studies

요 약

XML 문서에 대한 검색은 문서내 색인어 발생 빈도에 의한 내용 검색과 문서내 특정 엘리먼트에 의한 구조 검색 그리고 내용과 구조를 모두 검색하는 혼합 검색 등이 있다. 본 논문에서는 사용자의 질의에 대하여 문서에 의존하는 것이 아니라, 링크를 검색하여 특정 색인어에 대하여 가장 많은 링크를 설정 받은 문서 또는 특정 엘리먼트를 검색하는 새로운 검색 시스템을 설계한다. 이를 위해 XML 문서를 저장할 때 구조 정보와 함께 링크 정보를 저장하고 XML 링크에서 의미 속성인 ROLE, TITLE을 색인한다. 제안된 색인 모델에서는 정보를 찾는 사용자들의 질의를 보다 다양한 시각에서 검색할 수 있으며, 따라서 이러한 사용자들의 질의 유형과 그 처리 과정을 설명하고 의미를 분석한다.

1. 서론

웹과 정보 기술이 급속하게 발전함에 따라 필요한 정보를 찾기 위해 거의 대부분의 사람들이 웹에 의존하고 있다. 이에 정보를 찾으려는 사용자는 보다 양질의 정보를 정보 검색 시스템에 요구하고 있으며, 이러한 사용자의 요구를 충족시키기 위해 보다 향상된 정보 검색 시스템에 대한 연구가 계속되어지고 있다. 따라서 이러한 요구에 발맞추어 정보 검색 시스템의 성능을 향상시키는 것 뿐만 아니라, 정보를 찾는 시각을 보다 다양화 할 필요가 있다.

웹 상의 정보 교환의 수단으로 보편화된 HTML은 제한된 표현력과 의미론의 부재로 인하여 자원의 검색과 교환시에 사람의 도움을 많이 필요로 한다. 이에 보다 효율적인 표현과 의미 있는 정보 교환의 수단으로 XML이 각광을 받고 있다. DTD를 기반으로 한 XML 문서는 문서의 내용과 의미를 가지는 구조 정보, 그리고 문서와 문서, 문서와 특정 엘리먼트 사이의 링크 정보를 가지고 있다. 이 중에서 구조 정보는 문서가 내포하는 정보 관리를 보다 효율적으로 수행할 수 있을 뿐만 아니라, 정보를 교환함에 있어서도 정보의 내용에 의미를 부여한 교환이 가능하다. 따라서 이러한 구조 정보를 이용한 검색 시스템에 대한 연구가 활발히 진행되고 있다. 또한 XML 링크 정보는 그 기능을 더욱 발전시켜서 보다 다양한 역할을 수행한다. 따라서 문서와 문서 또는 문서와 특정 엘리먼트 사이의 관계가 다양하게 정의될 수 있기 때문에, 이러한 문서사이의 관계에 바탕을 둔 검색 시스템의 개발이 필요하다. 이전의 연구[2]에서 XML 링크를 기반으로 한 링크 검색 시스템에 대하여 연구한 바 있다. [2]에서 XML 링크를 이용할 시점의 XML 표준은 링크의 의미 속성인 ROLE과 TITLE 속성값을 위한 어떤 "승인된" 값을 미리 정의해 두지 않았었다. 또한 [2]에서는 XML 문서의 구조 정보를 무시하고 연구했기 때문에 XML 링크중에서 XPointer를 고려하지 않았다.

본 논문에서는 [2]의 논문을 개선하여 XLinks와

XPointer를 모두 적용한다. 이를 위해 XML 문서를 저장함에 있어 문서의 구조 정보와 함께 문서내 링크 정보를 저장한다. 그리고 문서를 색인함에 있어 내용에 기반한 색인과 구조에 기반한 색인, 그리고 링크 정보에 기반한 색인을 함께 한다. 여기서는 사용자의 질의에 대하여 링크 자체를 검색하기 위한 링크 기반 색인만을 다루고자 한다. 제안된 모델은 사용자의 질의어에 대하여 링크 검색을 가능하게 함으로써 정보를 찾는 시각을 보다 다양화할 수 있을 뿐만 아니라, 보다 양질의 정보를 찾는 사용자들의 욕구를 다양하게 충족시킬 수 있다.

2. XML 링크

2.1 링크 식별자 정의

XML 링킹 메커니즘은 내부 작업을 다루는 두가지 고유 스펙으로 XLink와 XPointer가 있다[3,4]. XLink는 이전의 XLL(eXtensible Linking Language) XML 문서가 또 다른 문서에 링크되는 방식을 세부적으로 기술하는 언어이다. XPointer는 링크가 문서 안의 다양한 장소를 가리키는 방식을 세부적으로 지정한다. XLink는 정의된 속성별로 다양한 종류가 있다. 이전의 연구[1]에서 XML 링크의 속성 중 유형의 정의하는 TYPE 속성과 행위를 정의하는 ACTUATE, SHOW 속성을 이용하여 각 링크의 식별자(ID)를 정의하였다(표 1).

<표 1> 링크 식별자 테이블

링크 속성			식별자
TYPE	ACTUATE	SHOW	
SIMPLE/ Inline_Extended	onLoad	Embed	1
	onLoad	Replaced	2
	onLoad	New	3
	onRequest	Embed	4
	onRequest	Replaced	5
	onRequest	new	6

2.2 링크의 의미(Semantics) 속성

XML 링크에는 의미(Semantics)와 관련된 속성, ROLE과 TITLE이 있다. 링킹 엘리먼트는 링크가 지시하는 원격 문서나 원격 문서의 특정 엘리먼트에 대한 추가적인 설명을 하기 위하여 이 속성들을 설정할 수 있다. TITLE 속성은 일반적으로 원격 문서의 내용을 설명할 수 있는 보통의 쉬운 텍스트를 값으로 가진다. ROLE 속성은 원격 문서를 보다 자세하게 설명하고 있는 문서에 대한 URI를 값으로 가진다. 따라서 이러한 TITLE과 ROLE 속성 값은 해당 링크가 가리키는 원격 문서에 대한 메타 데이터로서 작용을 한다.

2.3 설계 방향

본 논문에서는 XML 링크의 의미 속성을 기반으로 한 정보 검색 시스템을 제안함에 있어서, 사용자의 질의에 대하여 문서의 내용을 검색하는 것이 아니라 링크를 검색하여 문서의 우선 순위를 부여한다. 이를 위해 XML 문서를 저장할 때, 문서의 구조 정보와 함께 링크 정보 즉, 링크를 포함하는 엘리먼트와 원격 문서에 대한 주소 그리고 링크의 TITLE, ROLE 값을 저장한다.

문서와 문서사이의 링크를 기반으로 한 검색을 위해 색인 시, 링크의 의미 속성인 TITLE과 ROLE 속성을 각각 색인한다. TITLE 속성 값을 색인하여 특정 색인어로서 incoming 링크를 가진 문서를 색인한다. TITLE 과 ROLE 속성 값에 있는 용어로서 incoming 링크를 많이 가진 문서는 해당 용어로서 일반적으로 또는 보다 상세하게 설명되어질 수 있는 문서이다. 그리고 incoming 링크를 가진 문서를 색인하여, 해당 문서로 incoming되는 링크를 검색하고 해당 문서를 일반적으로 설명하고 있는 TITLE 값을 검색할 수 있게 하고, 또한 해당 문서를 보다 자세하게 설명하고 있는 ROLE 값이 지시하는 설명 문서를 검색할 수 있게 한다.

3. XML 문서 저장 구조 기법

XML 문서의 저장을 위해 먼저, DTD에 대한 구조 정보를 저장하고 모든 문서 인스턴스를 저장한다. 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.

하나의 XML 문서에 있는 모든 정보를 효율적으로 관리하기 위하여, 문서 인스턴스 정보를 구성함에 있어 엘리먼트 구조 정보, 링크 정보, 애트리뷰트 정보로서 구성한다(그림 1). 문서 인스턴스 정보는 문서를 구별하기 위한 식별자(DOC_ID), 문서안의 각

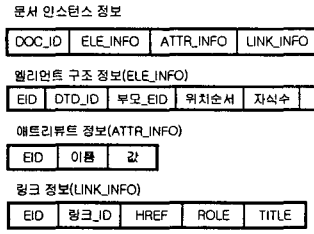

그림 1 XML 문서 정보 테이블

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.

링크 정보 테이블은 EID, 링크_ID, HREF, ROLE, TITLE로 구성된다. EID는 어느 엘리먼트에 속하는 링크인지를 식별하기 위하여 문서내 링크를 포함하는 엘리먼트의 EID를 저장한다. 링크_ID는 표 1에서 분류한 링크 식별자이다. 이 링크 식별자를 이용하여 링크_ID에 해당 링크를 분류하여 저장한다. HREF는 해당 링크가 지시하는 원격 문서 또는 원격 문서의 특정 엘리먼트의 주소이다. ROLE은 링크가 지시하는 원격 문서를 보다 자세하게 설명하고 있는 문서에 대한 URI 이다. TITLE은 링크가 지시하는 원격 문서를 일반적으로 쉽게 설명하고 있는 텍스트이다.

4. 링크 의미 기반의 정보 검색 시스템

4.1 색인 구조

여기서는 3장에서 구성한 XML 문서의 저장 구조 테이블 중에서 링크 정보를 이용하여 사용자의 질의어에 대하여 문서간의 링크를 보다 효율적으로 검색할 수 있는 링크 색인 구조를 설계한다.

먼저, 링크의 ROLE 속성값에 대한 색인 테이블은 ROLE_색인어 파일, ROLE_포스팅 파일, inlinks문서 파일, inlinks엘리먼트 파일로 구성된다(그림 2). ROLE_색인어 파일은 색인어와 노드수로 이루어진다. 색인어는 링크가 가지는 ROLE 속성값이 지시하는 문서에 나타나는 특정 용어이다. 노드수는 특정

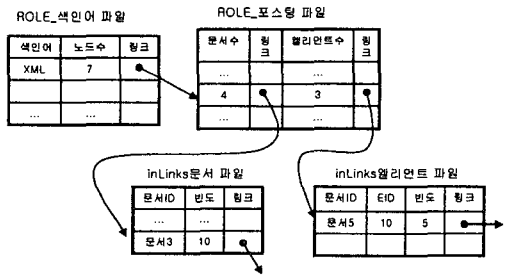

그림 2 ROLE 색인 파일 구조

색인어를 상세 설명 용어로 포함하고 있는 incoming 링크를 가지는 문서와 엘리먼트의 개수이다. ROLE_포스팅 파일은 특정 색인어를 ROLE 값에 포함하고 있는 문서의 수와 엘리먼트의 수를 나누어 저장하고 각각의 문서와 엘리먼트를 지시하고 있다. inlinks문서 파일은 문서ID, 빈도로 이루어진다. 문서ID는 incoming 링크를 가지는 문서의 식별자이고, 빈도는 해당 색인어로서 해당 문서가 incoming링크를 몇 개 가지는지에 대한 개수이다. inlinks엘리먼트는 문서ID, EID, 빈도로 구성된다. 문서ID는 incoming 링크를 가지는 문서의 식별자이고, EID는 해당 문서에서 링크가 지시하는 엘리먼트의 식별자이다. 빈도는 해당 엘리먼트가 해당 색인어로서 incoming 링크를 몇 개 가지는지에 대한 개수이다. 이러한 ROLE 색인 파일 구조로서 특정 색인어로서 가장 많은 incoming 링크를 가지는 문서나 특정 엘리먼트를 검색할 수 있다.

링크의 TITLE 속성값에 대한 색인 테이블도 ROLE 속성값에 대한 색인 테이블과 같다. 문서를 작성하는 사람은 특정 문서에 대한 링크를 설정할 때, 원격 문서가 특별한 경우가 아니라면 ROLE 속성값과 TITLE 속성값을 항상 설정하지 않을 것이다. 두 개의 속성 중에서 하나의 속성값이라도 가진다면 링크가 지시하는 원격 문서에 대하여 어떠한 목적으로 링크를 설정했는지 파악할 수 있다. 따라서 두 개의 속성값에 대하여 모두 색인할 필요가 있다.

다음으로, incoming링크를 가지는 문서에 대한 색인 구조이다(그림 3). incoming링크를 가지는 각각의

문서에 대해서 보다 효율적인 검색을 행하기 위하여 inLinks 문서 색인 테이블, inLinks문서 포스팅 파일로 구성한다. inLinks 문서 색인 테이블은 문서(엘리먼트), 횟수, ROLE문서로 구성된다. 문서(엘리먼트)는 incoming 링크를 가지는 문서나 엘리먼트에 대한 식별자이다. 횟수는 해당 문서나 엘리먼트에 대한 incoming 링크의 개수이다. ROLE 문서는 해당 문서나 엘리먼트의 incoming링크중에서 ROLE 속성값을 가진 링크가 있다면, 해당 ROLE 속성값이 지시하는 원격 문서의 개수이다. 이 개수는 해당 문서나 엘리먼트를 보다 자세하게 설명하고 있는 문서의 개수라고 볼 수 있다. inLinks문서 포스팅 파일은 HREF와 TYPE으로 구성된다. HREF는 해당 문서에 대한 보다 자세한 설명을 하고 있는 문서의 주소이다. TYPE은 지시하는 문서가 HTML 문서인지, XML 문서인지, 일반 텍스트 문서인지에 대한 분류이다.

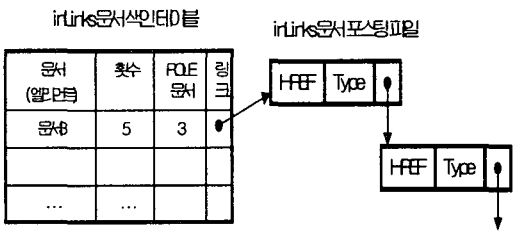

그림 3 incoming링크를 가지는 문서 색인 구조

4.2 질의어 유형 및 검색 과정

위의 색인 구조를 기반으로 링크의 메타 데이터를 검색하기 위한 질의어 유형을 가정하고, 각각의 유형별 검색 과정을 설명한다.

유형 1 : role/title = 'xml' and document(element)

'xml' 이라는 role 또는 title 메타 데이타 값으로 가장 많은 incoming 링크를 가진 문서(엘리먼트)를 빈도에 따라 우선 순위를 부여한다.

① ROLE(TITLE)_색인어 파일에서 'XML'이라는 색인어를 검색한다. ② 색인어가 지시하는 ROLE(TITLE)_포스팅 파일에서 해당 문서(엘리먼트)를 선택한다. ③ 문서(엘리먼트)의 링크가 지시하는 inLinks문서(inLinks엘리먼트) 파일에 있는 문서(엘리먼트)에서 빈도가 높은 순으로 우선 순위를 부여한다.

유형 2 : role/title = 'xml'

'xml' 이라는 role 또는 title 메타 데이타 값으로 가장 많은 incoming 링크를 가진 문서와 엘리먼트를 모두 검색하여 빈도에 따라 우선 순위를 부여한다.

유형 3 : inLinks = document(element)

특정 문서(엘리먼트)의 incoming 링크의 개수를 검색하라.

① inLinks문서(엘리먼트) 색인 테이블에서 해당 문서(엘리먼트)를 검색한다. ② 검색된 문서(엘리먼트)의 incoming 링크의 횟수를 출력한다.

유형 4 : inLinks = document(element) and value=role

특정 문서(엘리먼트)의 incoming링크의 role 속성에서 지시하는 문서를 검색하라. 즉, 해당 문서(엘리먼트)를 보다 자세하게 설명하고 있는 문서를 검색하라.

① inLinks문서(엘리먼트) 색인 테이블에서 해당 문서(엘리먼트)를 검색한다. ② 검색된 문서(엘리먼트)의 ROLE 문서값이 0 가 아니면 ③ 지시하는 inLinks문서 포스팅 파일의 href를 검색하여 나열한다.

5. 결론

본 논문에서는 사용자의 질의어에 대해 문서의 내용을 검색하는 것이 아니라 링크를 검색하는 새로운 정보 검색 모델을 제시하였다. 이를 위해 XML 링크의 의미(Semantics) 속성인 ROLE, TITLE 속성값을 색인하여 특정 용어에 대하여 incoming 링크를 가지는 문서나 엘리먼트 검색을 가능하게 하였다. 또한 incoming 링크를 가지는 문서나 엘리먼트를 색인하여 특정 문서나 엘리먼트가 가지는 incoming 링크를 검색하고, 해당 문서나 엘리먼트를 보다 자세하게 설명하고 있는 role이 지시하는 문서의 검색을 가능하게 한다. 제안된 모델은 정보 검색 시스템으로부터 보다 양질의 정보를 찾기 원하는 사용자에게 보다 다양한 시각에서의 정보 검색을 제공할 수 있다.

참고문헌

[1] 김은정, 배종민 "XLinks를 이용한 하이퍼텍스트 검색 시스템", 한국정보처리학회논문지, 제 8권 5호, p483~494, 2001.

[2] 김상준, 김은정, 배종민 "XML 링크의 메타데이타를 이용한 검색 시스템의 설계", 한국정보과학회봄학술발표논문집, Vol.27,No.1.,p157~159, 2000

[3] W3C Recommendation 27-June-2001, "XML Linking Language(XLink)", http://www.w3.org/TR /xlink

[4] W3C Working Draft 16-August-2002, "XML Pointer Language(XPointer)", http://www.w3.org/TR/xptr