XML 링크의 메타데이타를 이용한 검색 시스템의 설계

김 상 준°, 김 은 정, 배 종 민
경상대학교 컴퓨터과학과 / 정보통신연구센터

Design of a Retrieval System using Metadata in XML Links

Sang-Joon Kim°, Eun-Jung Kim, Jong-Min Bae
Dept. of Computer Science / Information and Communication Research Center
Gyeongsang National University

요 약

인터넷의 보편화로 정보 교환이 활발해지면서 일반 사용자들에게 필요한 정보를 손쉽게 취득하게 해주는 정보 검색 시스템의 역할이 아주 중요하게 되었다. 일반적인 정보 검색은 사용자의 질의에 대해 문서내의 색인어 발생 빈도를 기반으로 관련 문서를 찾아준다. 본 논문에서는 XML 링크 정보를 이용하여 링크를 검색하여 관련 문서를 찾아주는 정보 검색 시스템을 제시한다. 이를 위해 XML 링크에서 ROLE 속성값인 메타 데이터를 색인하여 저장하고 링크에 대한 모든 정보를 저장하고, 이를 기반으로 특정 주제에 대한 검색시, 특정 주제로 가장 많이 링크된 문서를 검색한다. 제시한 방법을 현재 웹상에서 주로 이용되는 HTML 문서를 기반으로 분석해 본 결과 그 필요성을 확인할 수 있었다.

1. 서론

최근 웹과 정보 기술의 급속한 발전에 따라 개개인이 소화해야 할 정보의 양도 그만큼 많이 늘어나고 있다. 또한 하나의 사용자 질의에 대한 정보 검색 시스템들의 용답이 모두 다를 뿐만 아니라 질의에 대한 결과로서 많은 문서가 출력되기 때문에 출력된 문서중에서 필요한 것을 찾기가 점점 어려워지고 있다. 이로 인해 오늘날 인터넷에서 정보를 찾으려는 일반 사용자들의 부담은 점점 커지고 있다.

일반적인 정보 검색 시스템은 사용자의 질의에 대해 문서 단위의 검색을 하며, 주로 문서내 색인어 발생 빈도를 문서의 순위를 부여하는데 이용한다. 이는 많은 검색 결과중에서 양질의 정보를 찾는데 한계가 있으며, 결국 질의에 대한 가장 적합한 정보를 찾는 것은 사용자의 몫이다. 따라서 정보를 찾는 시각을 보다 다양화할 필요가 있다.

HTML 문서는 한정된 태그 집합을 가지고 있어서 다양한 네트워 자원을 효율적으로 교환 및 검색하기에는 한계가 있다. 이에 대한 해결방안으로서 차세대 웹 언어로 제시된 것이 XML이다[1,2,3,4]. 특히 XML의 링크는 그 기능을 더욱 발전시켜서 보다 다양한 역할을 수행하기 때문에 검색에 있어서도 유용하게 이용되어질 수 있다.

본 논문에서는 XML 링크의 ROLE 속성을 이용하여 사용자의 질의에 대해 문서간의 링크를 검색하여 특정 주제로서 가장 많은 링크를 설정받은 문서순으로 순위를 부여한다. 또한 XML 링크의 다른 속성값을 이용하여 문서사이의 특정 관계에 기반한 검색을 가능하게 한다. 이를 위해 XML 링크의 여러 속성값에 따라 고유한 식별자를 부여하고, 문서를 색인할 때 문서안의 모든 링크에 대해 ROLE 속성 값을 색인하여 저장하고, 하나의 문서에서 나가는 링크 즉, OUTGOING 링크와 하나의 문서안으로 들어오는 링크 즉, INCOMING 링크에 대한 정보를 모두 저장하였다. 이에 본 논문에서는 이러한 검색 시스템을 위한 색인 구조와 이를 기반으로 하는 검색 질의어 유형 및 검색 과정을 제시한다.

2. XML 링크

2.1 링크 식별자 테이블 정의

링크는 정의된 속성별로 다양한 종류가 있다. 링크의 속성 중 TYPE, SHOW, ACTUATE 세 가지를 이용하여 각 링크의 식별자(ID)를 <표 1>과 같이 정의한다.

<표 1> 링크 식별자 테이블

TYPE	ACTUATE	SHOW	ID	TYPE	ACTUATE	SHOW	ID
simple	auto	parsed	1	extend ded	auto	parsed	7
	auto	replaced	2		auto	replaced	8
	auto	new	3		auto	new	9
	user	parsed	4		user	parsed	10
	user	new	5		user	new	11
	user	replaced	6		user	replace	12

2.2 링크의 ROLE 속성

XML 링크에는 의미(semantic)와 관련된 속성, ROLE이 있다. 이 속성은 일반적으로 링크된 원격 문서의 역할을 설명하기 위하여 사용되며, 속성값은 일반적으로 원격 문서의 내용을 총칭하는 스트링(string)이다. 즉, 링크의 ROLE 속성값은 해당 링크가 가리키는 원격 문서에 대한 메타 데이터로서 작용을 한다.

XML 표준은 링크의 ROLE 속성값을 위한 어떤 "승인된" 값을 미리 정의해 두지 않았다. 따라서 하나의 문서에 설정된 모든 INCOMING 링크들은 해당 링크를 OUTGOING 링크로 설정한 문서 작성자의 주관적인 생

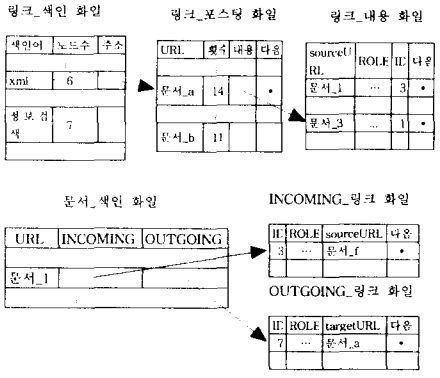
각에 따라 서로 다른 값을 가진다. 이는 하나의 문서에 대한 정확한 메타 데이터를 이해하는데 많은 어려움을 가진다. 따라서 본 논문에서는 하나의 링크를 설정할 때 ROLE 속성값을 부여함에 있어, 링크가 가리키는 원격 문서의 내용을 대표할 수 있는 단어들의 리스트를 콤마(,)로서 연결하는 방법으로 원격 문서에 대한 메타 데이터를 정의하는 것으로 가정한다. 예를 들면 다음과 같다.

<LINK href="문서_1" role="xml, html"> ... </LINK>
<LINK href="문서_1" role="정보검색"> ... </LINK>

3. 링크 기반의 정보 검색 시스템

3.1 색인 구조 및 과정

본 논문에서 제공하는 링크의 메타 데이터 검색을 위한 색인 구조는 링크_색인 화일, 링크_포스팅 화일, 링크_내용 화일, 문서_색인 화일, INCOMING_링크 화일, OUTGOING_링크 화일로 이루어진다. (그림 1)의 링크_색인 화일에는 링크의 메타 데이터를 대표하는 색인어들을 저장하고 링크_포스팅 화일에는 해당 색인어로서 링크된 원격 문서와 링크받은 횟수가 들어있다. 링크_내용 화일에는 링크_포스팅 화일에 있는 각 문서의 링크들의 속성 내용이 들어있다. 문서_색인 화일과 INCOMING_링크, OUTGOING_링크 화일은 특정 문서별로 문서의 모든 OUTGOING 링크와 INCOMING 링크에 대한 정보를 가지고 있다.



(그림 1) 링크 검색 시스템의 색인 화일 구조

하나의 문서에서 각 링크에 대한 색인과정은 다음과 같다.

① 문서_색인 화일에 해당 문서를 검색한 후, 등록한다.

② 문서에서 링크를 만나면, 링크의 메타 데이터에서 색인어를 찾아 링크_색인 파일에서 해당 색인어가 존재하면 (2.1), 존재하지 않으면 (2.2)를 수행한다.

(2.1) 해당 색인어가 가리키는 링크_포스팅 파일에 링크의 원격 문서가 존재하면 (2.1.1), 존재하지 않으면 (2.1.2)을 수행한다.

(2.1.1) 해당 문서의 횟수를 1 증가시키고 링크의 내용을 링크_내용 파일에 등록한다.

(2.1.2) 링크의 원격 문서를 url 필드에 등록, 링크 내용을 링크_내용 파일에 등록한 후, 링크_색인 파일의 해당 색인어의 노드수를 1 증가시킨다.

(2.2) 색인어를 링크_색인 파일에 등록하고, 노드수에 1 저장, 해당 링크에 대한 정보를 링크_포스팅 파일과 링크_내용 파일에 등록한다.

③ 문서_색인 파일의 해당 문서가 가리키는 OUTGOING_링크 파일에 링크를 등록하고, 링크의 원격 문서를 문서_색인 파일에서 찾아서, 존재하지 않으면 등록 후, 해당 문서의 INCOMING_링크 파일에 링크를 등록한다.

④ 문서의 끝까지 모든 링크를 찾아서 ① ~ ③의 과정을 반복한다.

3.2 질의어 유형 및 데이터 검색 과정

3.2.1 질의어 유형

위의 색인 구조를 기반으로 링크의 메타 데이터 검색을 하기 위하여 다음과 같은 질의어 유형을 가정한다.

유형 1 : search_link = (메타 데이터)
주어진 메타 데이터로서 INCOMING 링크가 가장 많이 설정된 문서 순위로 검색하라.

유형 2 : search_link = (메타 데이터) and show=(parsed/new/replaced) and actuate=(auto/user)
주어진 메타 데이터로서 INCOMING 링크가 설정된 문서중에서 지역 문서와 원격 문서간의 특별한 관계를 가지는 문서만을 검색하라.

유형 3 : search_link = (메타 데이터) and document = (문서의 URL) 특정 메타 데이터로서 특정 문서에 설정된 링크만을 검색하라.

유형 4 : document = (문서의 URL) and direction=(incoming/outgoing) 특정 문서의 incoming 또는 outgoing 링크에 대해서 모두 검색하라.

3.2.2 데이터 검색 과정

각각의 질의어 유형별 검색과정을 차례대로 나타내면 아래와 같다.

유형 1 : search_link= (xml)
① 링크_색인 화일에서 색인어 검색
② 링크_포스팅 화일에서 해당 문서들 출력

유형 2 : search_link = (xml) and show= (parsed) and actuate=(auto)
① 링크_색인 화일에서 'xml' 색인어 검색
② 링크_포스팅 화일의 내용 필드가 가리키는 링크_내용 화일을 검색
③ 자동 삽입되는 링크만 검색한 후 출력

유형 3 : search_link=(xml) and document= (문서_2)
① 링크_색인 화일에서 'xml' 색인어 검색
② 링크_포스팅 화일에서 url='문서_2'를 검색
③ 링크_내용 파일의 링크 내용 출력

유형 4 : document=(문서_2) and direction =(outgoing)
① 문서_색인 화일에서 '문서_2'를 검색
② outgoing 필드가 가리키는 내용을 출력

4. 실험 및 분석

여기서는 웹상에 존재하는 정보 검색기중에서 'Yahoo', 'Altavista'를 이용한 실험을 통하여 결과를 분석하였다. 검색기의 대상이 주로 HTML 문서이므로 HTML 문서를 이용하였다. 이를 위해 HTML 문서에 있는 링크의 앵커(anchor)를 해당 링크의 메타 데이터로 가정하였고 또한 모든 링크의 앵커가 질의어에 해당되는 것으로 가정하였다. 분석을 위하여 위의 2개의 정보 검색기를 대상으로 'xml'이라는 질의어로서 검색을 하였다. 검색 결과 각 상위 10개의 순위에 해당되는 문서의 URL을 살펴보면 <표 2>와 같다.

<표 2> Yahoo와 Altavista 검색기에서 'xml' 질의어에 대한 검색 결과

순위	Yahoo	Altavista
①	http://www.xml.com/	http://www.xml.com
②	http://www.xml.org/	http://www.oasis-open.org/cover/sgml-xml.html
③	http://www.ibm.com/developer/xml/	http://xml.com/
④	http://metalab.unc.edu/xml/	http://www.gca.org/conf/xml/xml-what.html
⑤	news: comp.text.xml	http://www.alphaworks-ibm.com/formula/xml
⑥	http://www.w3.org/xml/	http://www.xmlinfo.com/
⑦	http://www.geovities.com/	http://webreview.com/wr/pub/XML/
⑧	http://metalab.unc.edu/pub/sun-info	http://www.jclark.com/xml/canonxml.html
⑨	http://www.sciam.com/	http://www.textuality.com/sgml-erb/WD-xml.html
⑩	http://www.oasis-open.org/cover/	http://msdn.microsoft.com/

위의 2개의 검색 결과에는 상관없이 일반적으로 XML에 대한 좋은 정보를 가진것으로 많이 알려진 사이트들을 살펴보면 <표 3>과 같다. 그러나 <표 3>의 사이트들은 위의 2개의 검색기들의 검색 결과에서는 좋은 반응을 보이지 않았으며 어떤 사이트는 순위에 포함되지도 않았다.

<표 3> 많이 알려진 XML 사이트들

Top XML Sites
IBM's XML Web Site [www.ibm.com/developer/xml]
XML Developer Center[msdn.microsoft.com/xml/]
OASIS home page [www.oasis-open.org]
XML.org [www.xml.org]
The World Wide Web Consortium's XML [www.w3.org/xml/]
XML.COM [www.xml.com/xml/pub]
The XML Working Group FAQ [www.ucc.ie/xml/]

다음으로 본 논문에서 제시한 링크 검색 시스템의 검색 방법을 적용하기 위하여 우선 웹상에서 위의 각 상위 10개의 문서들 중에서 중복되는 것을 제외한 17개와 또다른 XML에 대한 문서들을 랜덤하게 골라 대략 100여개의 문서들을 조사하였다. 분석 방법은 각 문서의 OUTGOING 링크를 분석하여 링크가 가리키는 문서가 위의 17개의 사이트들중 하나인것만 조사하였다. 여기서 모든 링크는 메타 데이터로서 'xml'을 가지는 것으로 가정하였다. 따라서 100여개의 문서에 대한 조사가 끝나면 위의 17개의 사이트에 대해서 100여개의 문서로부터 몇번의 INCOMING 링크가 설정되었는지를 판단할 수 있다. 분석 결과는 <표 4>와 같다.

<표 4>의 결과에서 알수 있듯이 [http://www.w3.org/xml/] 사이트가 'xml'이라는 주제어로 가장 많은 INCOMING 링크를 가졌다. 이는 많은 사람들이 'xml'에 대한 정보를 얻기 위해 이 사이트를 참조한다

<표 4> 각 사이트의 링크받은 횟수

사이트	횟수	순위
http://www.xml.com/	39	②
http://www.xml.org/	23	③
http://www.ibm.com/developer/xml/	19	⑤
http://metalab.unc.edu/xml/	2	⑩
news: comp.text.xml		
http://www.w3.org/xml/	48	①
http://www.geovities.com/		
http://www.sciam.com/		
http://www.oasis-open.org/	9	⑦
http://xml.com/	20	④
http://www.gca.org/conf/xml/xml-what.html		
http://www.alphaworks-ibm.com/formula/xml		
http://www.xmlinfo.com/	3	⑧
http://webreview.com/wr/pub/XML/		
http://www.jclark.com/xml/canonxml.html		
http://www.textuality.com/sgml-erb/WD-xml.html	2	⑨
http://msdn.microsoft.com/	11	⑥

고 생각할 수 있다. 그러나 이 사이트는 위의 yahoo 검색기의 검색 결과에서는 낮은 순위를 보였을 뿐만 아니라 altavista 검색기의 검색 결과에서는 순위에 포함되지도 않았다.

결과적으로 보면, 정보 검색 결과는 질의를 던진 사용자의 주관적 판단외에는 그렇게 큰 의미를 부여할 수 없다. 따라서 사용자의 질의에 가장 가까운 문서를 찾는 것은 결국에는 사용자의 판단에 의해서 이루어진다. 때문에 보다 다양한 방법으로 사용자의 욕구를 충족해 줄수 있어야 한다. 이에 본 논문에서의 링크 검색 방법은 정보를 찾는 사용자에게 보다 다른 시각에서 원하는 정보에 접근할 수 있는 새로운 방법을 제공하였으며 위의 결과에서 그 필요성이 충분하다고 판단된다.

5. 결론 및 향후과제

본 논문에서는 차세대 웹 언어로 등장한 XML의 링크 정보를 이용하여 링크를 검색하여 원하는 정보를 검색하는 시스템을 제시하였다. 제시된 정보 검색기는 기존의 정보 검색기들의 검색 결과와는 다른 검색 결과를 볼 수 있으며, 이는 사용자에게 보다 다양한 시각에서의 정보 검색을 제공할 수 있다.

향후 연구과제로는 대규모 XML 문서 컬렉션을 대상으로 검색기의 구현 및 효율성을 검증하는 것이다.

참 고 문 헌

[1] Rohit Khare, Adam Rifkin "XML: A door to *Automated* Web *Applications*", IEEE Internet Computing, pp.78-87, July & August 1997.

[2] Ronald96, Ronald C. Tumothy A. Douglass, Audrey J. Turner "Readme.1st SGML for Writers and Editors", PH, 1996.

[3] W3C Working Draft 26 July 1999, " XML Linking Language(XLink)", http://www.w3.org/TR/ xlink

[4] W3C Working Draft 9 July 1999, "XML Pointer Language(XPointer)",http://www.w3.org/TR/WD-xptr