XML 링크정보를 이용한 정보 검색 색인 기법의 설계

김 은 정† · 배 종 민††

요 약

웹 환경에서 정보 교환의 주된 표현으로서 하이퍼텍스트 문서가 이용되고 있다. 하이퍼텍스트 문서의 구조는 링크로 서로 연결된 그래프의 형태로서 문서를 비순차적으로 처리할 수 있게 한다. 본 논문에서는 XML 링크 정보를 이용한 정보검색시스템을 위한 색인 기법을 제시한다. 이를 위해 XML 문서에서 링크된 원격 문서에 있는 링크들을 제어할 수 있는 새로운 속성을 정의하고, 각 링크들에 대해 속성별로 고유한 식별자를 부여한다. 식별자는 지역 문서와 원격 문서에서 서로 다른 가중치를 가지며, 이러한 가중치를 기반으로 하나의 문서를 색인할 때 지역 문서뿐만 아니라 링크가 가리키는 원격 문서에서도 색인을 한다. 이러한 방법은 문서에서의 링크에 대한 정보를 무시하고 각 문서를 독립적으로 색인한 기존의 방법에 비하여 사용자의 요구사항에 대하여 보다 근접한 문서를 검색할 수 있다.

Design of an Information Retrieval Indexing Method using XML Links

Eun-Jung Kim† · Jong-Min Bae††

ABSTRACT

The hypertext document is used for information exchange in the Web environments. Its structure is considered as having graph structures with links, which makes nonlinear processing of documents possible. This paper proposes an indexing method for information retrieval system using XML links. We define new attributes that control links of a remote document and assign an unique identifier for the attribute of each link. Each identifier has a different weight according to its occurrence position that is local or remote documents. We index a word not only from a local document but a remote document based on the given weight. Experimental results show that the proposed method outperforms conventional retrieval systems that ignore links.

1. 서 론

인터넷의 보편화로 정보 교환이 활발해지면서 필요한 정보를 손쉽게 취득하게 해 주는 정보 검색 시스템의 역할이 과거 어느 때보다도 중요하게 되었다. 일반적인 정보 검색은 사용자의 질의에 대해 정형 혹은 비정형 자료를 문서 단위로 찾아주는 기능을 말하며, 근래에는 인터넷 자료만을 전문적으로 검색하는 시스템이 많이 보급되고 있다[6, 14]. 이러한 검색 시스템은 주로 문서를 검색 단위로 간주하고 있는데, 문서와 문서 사이의 관계에 대한 정보는 검색 대상에서 제외된다.

일반적으로 대부분의 문서는 묵시적이든 명시적이든 다양한 구조 정보를 가지고 있다. 문서의 장, 절, 제목, 참고문헌 등을 명시하는 논리적인 구조 정보와 타 문서와의 연계성 구조 정보, 그리고 문자형, 쪽 구분 등

† 준 회 원 : 경상대학교 대학원 전자계산학과
†† 종신회원 : 경상대학교 컴퓨터과학과 교수
논문접수 : 1999년 11월 18일, 심사완료 : 2000년 6월 3일

제시(presentation)용 정보로 나눌 수 있다. 특히, 타 문서간의 연계성 구조 정보는 하이퍼텍스트 형태로서 명시적으로 표현될 수 있는데, 이는 문서를 선형적으로 보지 않고 작은 노드라고 불리는 텍스트 조각이 링크로 서로 연결되어 그래프를 형성하고 있는 형태로써 문서 처리를 비선형적 혹은 비순차적으로 할 수 있게 한다. 현재 웹 환경에서 보편화된 HTML 문서도 하이퍼텍스트 형태로 정보검색의 대상이 되어 왔으나 대부분의 웹 검색 엔진의 경우 링크 정보를 무시하고 각 노드를 독립적인 문서로 간주하여 검색한다.

HTML 문서는 한정된 태그 집합을 가지고 있어서 다양한 네트웍 자원을 효율적으로 교환 및 검색하기에는 한계가 있다. 또한, 제한된 표현력과 의미론의 부재로 인하여, 자원의 검색과 교환시에 사람의 도움을 많이 필요로 한다. 이에 대한 해결방안으로서 차세대 웹 언어로 제시된 것이 XML(eXtensible Markup Language)이다[12, 13]. XML은 W3C(World Wide Web Consortium)에 의해서 개발된 ISO의 SGML을 인터넷에서 사용할 수 있도록 단순화한 버전이다[7-9]. XML의 문서구조는 논리적 구조와 물리적 구조로 구성된다. 논리적 구조는 문서의 전체적인 구조를 표현하는 부분으로 엘리먼트, 속성 등의 구성요소를 이용하여 표현하며, 물리적 구조는 엔티티를 이용하여 표현하며, 이들이 혼합되어 문서의 구조를 나타내는 DTD를 형성한다[9-11]. 특히 링크의 경우, HTML에서의 링크가 단순히 타 문서와의 연결을 설정해 주고 브라우징 시에 사용자가 선택하면 해당 문서로 항해할 수 있게 하지만, XML의 링크는 그 기능을 더욱 발전시켜서 보다 다양한 역할을 수행한다.

XML 링크에는 다양한 의미를 가지고 있어서, 한 문서와 그 문서에서 링크된 문서 사이의 관계가 다양하게 정의될 수 있기 때문에, 문서 사이의 관계에 바탕을 둔 검색시스템의 개발이 필요하다. 이를 위해 본 논문에서는 하이퍼텍스트 문서를 검색함에 있어서, XML 링크 정보를 이용한 정보검색시스템을 위한 색인 기법을 제시한다. 제시된 시스템에서는 한 문서에서 발생한 색인어를 색인할 때, 그 문서 즉, 지역문서 뿐 아니라 그 문서와 일정한 관계를 가진 문서 즉, 원격문서에 대해서도 고려한다. 이때 문서사이의 관계성에 따라서 관계성의 정도를 나타내는 가중치가 다르게 주어지며, 이 관계성의 정도를 나타내는 가중치에 바탕을 두어서 문서를 색인한다.

본 논문의 2장에서는 관련 연구 및 설계 방향을 제시하고 3장에서는 본 논문에서 제안하는 링크 정보를 이용한 정보 검색 시스템을 위한 전반적인 과정에 대하여 설명한다. 이를 바탕으로 4장에서 기존의 색인 방법과 본 논문에서 제시하는 링크 정보를 이용한 색인 방법을 실험하여 분석하고 평가한다. 마지막으로 5장에서 결론 및 향후과제를 보인다.

2. 관련 연구 및 설계 방향

2.1 XML 링크

XML 링킹 메커니즘은 내부 작업을 다루는 두 가지 고유 스펙으로 XLink와 XPointer가 있다[12, 13]. XLink는 이전의 XLL(eXtensible Linking Language)로 XML 문서가 또 다른 문서에 링크되는 방식을 세부적으로 기술하는 언어이다. XPointer는 링크가 문서 안의 다양한 장소를 가리키는 방식을 세부적으로 지정한다. XLink를 사용하면 문서를 연결할 수 있고, XPointer를 사용하면 문서 자체 안에 있는 특정 지점을 참조할 수 있다.

XLink에는 정의된 속성별로 다양한 종류가 있다. Xlink에서 정의할 수 있는 속성들의 종류는 다음과 같다.

① TYPE 속성 : XML에서는 링크를 위해서 엘리먼트 선언시 TYPE 속성을 선언하여 다양한 링크의 유형을 정의한다. 선언 가능한 값으로 simple, extended, locator, resource, arc, title 이다

② SHOW 속성 : SHOW 속성은 링크된 내용을 웹 브라우저에 디스플레이시 사용자에게 어떻게 보여줄 지에 대한 기능을 정의한다. 선언 가능한 속성값은 embed, replaced, new이다. embed는 문서가 디스플레이시 링크된 자원이 문서안에 포함되도록 한다. replaced는 문서가 디스플레이시 링크된 자원이 링크한 위치에 대체되어 나타나게 한다. new는 문서가 디스플레이시 링크된 자원이 새로운 윈도우나 프레임에 나타나도록 한다.

③ ACTUATE 속성 : ACTUATE 속성은 링크를 어떻게 활성화할 것인지에 대하여 정의한다 속성값으로는 onLoad와 onRequest가 있다. onRequest는 사용자가 링크를 선택했을 시에만 링크가 활성화된다. onLoad는 문서가 디스플레이 또는 처리시에 링크가 자동적으로 활성화된다.

④ ROLE 속성 : ROLE 속성은 일반적으로 링크된 원격 문서의 역할을 설명하기 위하여 사용되며, 속성

값은 XNAME에 정의된 QName으로서 Nmtoken 이다.

⑤ TITLE 속성 : 라벨로서 역할을 한다.

2.2 하이퍼텍스트 기반의 정보 검색 시스템

하이퍼텍스트 형태로 정보 검색의 대상이 되는 문서에 있어서 링크는 유용한 정보이다. 타 문서간의 연계성 구조 정보인 링크는 그 분류 방법에 따라 여러 가지로 나눌 수 있다. 하이퍼텍스트 기반의 정보 검색에 대한 연구가 많이 진행되어 왔다[2-5]. 이 중에서 링크를 분류함에 있어 링크의 방향성과 직접/간접 링크에 중점을 두고 하이퍼텍스트에 대해 링크 정보 적용 방법을 제시한 연구가 있다[5]. 링크의 방향성은 한 문서에서 다른 문서로 나가는 링크와 다른 문서에서 링크되어 들어오는 링크로 구분하고, 직접/간접 링크는 문서끼리 직접 연결한 링크와 다른 문서를 거쳐서 연결한 링크로 구분하였다. 링크 정보를 효과적으로 적용하기 위한 기본 접근 방법으로서 질의어와 문서간의 RSV(Retrieval Status Value)를 조정하도록 하였으며, RSV는 질의어와 문서간의 함수로서 링크의 앵커(anchor)가 질의어에 해당되는지 아닌지에 따라 링크 적용 효과를 달리 하도록 하였다 따라서 문서의 색인시 각 문서를 독립적으로 색인하고, 색인과정에서 링크 정보를 위의 분류 방법에 따라 분류, 저장하였다. 그리고 사용자의 질의에 대해 문서의 순위를 결정하는 과정에서 색인 결과에서 나온 문서의 집합에 링크 정보를 적용하여 문서 집합의 확장 및 순위를 재조정하였다

본 논문에서는 XML 링크 정보를 이용한 정보 검색 시스템을 제안함에 있어 링크 정보를 문서의 색인과정에서 이용한다. 제안하는 시스템에서는 하나의 문서를 색인하는 과정에서 지역 문서뿐만 아니라 문서내의 링크가 가리키고 있는 원격 문서도 색인을 한다 이를 위해 하나의 문서에 있는 링크가 자신이 가리키는 원격 문서에 있는 링크를 제어할 수 있는 새로운 속성을 정의하였다. 또한 각 속성별로 정의된 서로 다른 링크들에 대해 고유한 식별자를 부여하여 각 식별자별로 링크의 가중치를 부여하였다. 이러한 링크의 가중치는 해당 링크가 지역 문서에 있을 때와 원격 문서에 있을 때 그 값이 달라진다. 따라서 하나의 문서를 색인할 때, 링크된 원격 문서는 링크의 가중치 값을 기준으로 색인을 하며, 원격 문서 안의 링크들에 대해서는 그 속성에 따라 계속해서 색인할 지의 여부를 판단한다. 이 방법은 문서에서의 링크에 대한 구조 정보를 무시하고 각 문서를 독립적으로 간주하여 색인한 기존의 방법에 비해서 사용자의 질의에 보다 근접한 문서를 찾아낼 수 있다.

3. 링크 정보를 이용한 정보 검색 시스템

3.1 원격 문서의 링크에 대한 제어 속성 정의

XML 링크의 속성에는 링크된 원격 문서 안에 있는 링크에 대해서 지역 문서에서 제어할 수 있는 방법은 제공되지 않는다. 예를 들어, 지역 문서에 있는 링크들 중에서 속성이 show가 'embed'이고 actuate가 'onLoad'인 링크가 있다고 가정하자. 이 링크는 지역 문서가 활성화될 때 링크된 원격 문서의 내용이 자동적으로 지역 문서에 삽입되게 된다 이때 삽입된 원격 문서 안에 다시 똑같은 링크가 존재한다면 해당 링크도 자동적으로 문서 안에 삽입된다 사용자가 원격 문서 안에 있는 해당 링크는 지역 문서가 활성화될 때 자동적으로 삽입되기를 원하지 않더라도 XML 링크에는 원격 문서 안의 링크에 대한 제어 속성이 없기 때문에 사용자가 원격 문서 안의 링크에 대해서 활성화 여부를 제어할 수 없다 따라서 본 논문에서는 이런 경우 원격 문서 안의 어떤 링크에 대해서 지역 문서에서 해당 링크를 제어할 수 있는 새로운 속성을 정의한다. 정의하는 속성은 다음과 같다.

3.1.1 REMOTE_LINK

원격 문서 안의 링크를 그대로 포함시킬지의 여부를 정의하는 속성으로서 가질 수 있는 값은 YES와 NO이다.

- YES : 링크된 원격 문서 안의 모든 링크를 그대로 수용한다. 이 속성에 대한 정의가 생략된다면 디폴트값이 YES 이다.
- NO : 링크된 원격 문서 안의 링크에 대해 모두 수용하지 않는다. 이 경우, 수용할 링크만 INSERT-LINK 속성에서 정의한다.

3.1.2 INSERT_LINK

REMOTE_LINK의 값이 NO일 경우, 링크된 원격 문서 안에 있는 링크를 모두 수용하지 않는다. 이때 사용자가 수용하고자 하는 링크만 이 속성에서 정의한다.

이 속성이 가질 수 있는 값은 링크에 대한 ID 이다.

- ID Number : 본 논문에서는 링크의 종류에 따라 식별자(ID)를 부여한다 이 속성에서 각 링크에 대한 해당 식별자를 선택하여 열거한다. 이 속성에서 열거된 링크들만이 원격 문서 안에서 원래의 속성대로 활성화된다 링크 식별자에 대한 자세한 내용은 3.2절에서 설명한다.

3.2 링크 식별자 테이블 정의

문서내의 링크는 내용의 흐름상 지역 문서와 링크된 원격 문서가 얼마나 관련성이 있는가 하는 것이 문제이다. 링크는 정의된 속성별로 다양한 종류가 있다. 링크의 종류를 분류하기 위해 링크의 속성 중 TYPE, SHOW, ACTUATE 세 가지를 이용한다 각 속성 값에 따른 링크의 식별자(ID)를 <표 1>과 같이 정의한다.

〈표 1〉 링크 식별자 테이블 ($0 \leq \delta < \gamma < \beta < \alpha \leq 1$)

TYPE	ACTUATE	SHOW	ID	가중치	
				지역	원격
simple/ extended (inline경우)	onLoad	embed	1	α	α or 0
	onLoad	replaced	2	α	α or 0
	onLoad	new	3	α	α or 0
	onRequest	embed	4	β	β or 0
	onRequest	new	5	γ	γ or 0
	onRequest	replaced	6	δ	0

<표 1>에서의 분류는 XML 링크의 행동과 관련된 속성에 기준하여 단순 링크와 인라인 확장 링크만을 고려하였다. 각 ID에 따른 링크의 행동은 <표 2>와 같다.

〈표 2〉 각 ID에 대한 링크의 행동

ID	링크의 행동
1	자동적으로 원격 문서가 지역 문서에 삽입.
2	자동적으로 원격 문서가 대체되어 나타남.
3	새로운 윈도우나 프레임에 원격 문서가 자동적으로 나타남.
4	사용자의 선택에 의해 원격 문서가 지역 문서에 삽입.
5	사용자의 선택에 의해 새로운 윈도우나 프레임에 원격 문서가 나타남
6	사용자의 선택에 의해 원격 문서가 대체.

각 링크는 서로 다른 가중치값을 가지며 이 가중치 값은 원격 문서에 있는 용어의 가중치를 계산할 때 사용된다. 즉, 하나의 문서를 색인 함에 있어 지역 문서에서 나타나는 용어의 가중치를 1로 부여할 때 원격 문서에 나타나는 용어의 가중치는 해당 링크의 가중치 값으로 부여한다.

링크가 지시하는 원격 문서에는 또다시 링크가 존재할 수 있다. 이때 해당 링크가 지시하는 원격_원격 문서 안에서의 용어의 가중치 계산은 지역 문서에서 원격 문서를 지시한 링크의 속성에 따라 달라진다. 본 논문에서는 지역 문서에서 링크된 원격 문서 안에 있는 링크들에 대한 제어를 위해 새로운 속성 REMOTE_LINK와 INSERT_LINK를 정의하였다. <표 1>에서 정의한 각 식별자(ID)마다 해당 링크가 정의될 때 REMOTE_LINK와 INSERT_LINK의 속성 값에 따라 원격 문서내의 링크에 대한 가중치가 달라진다 원격 문서안의 링크에 대한 가중치 부여 방법은 <표 3>과 같다.

〈표 3〉 원격 문서안의 링크에 대한 가중치 부여

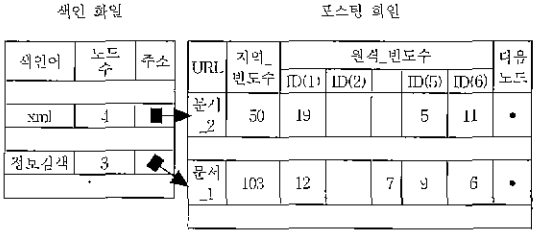
ID	지역문서의 링크 속성값	가중치
1,2,3	remote_link='yes' or insert_link='1,2,3' 포함	α
	remote_link='no' and insert_link='1,2,3' 포함 안됨	0
4	remote_link='yes' or insert_link='4' 포함	β
	remote_link='no' and insert_link='4' 포함 안됨	0
5	remote_link='yes' or insert_link='5' 포함	γ
	remote_link='no' and insert_link='5' 포함 안됨	0
6	두 개의 속성값에 상관안함	0

원격 문서안에 있는 링크가 지시하는 원격_원격 문서는 지역 문서의 링크가 가지는 remote_link와 insert_link의 속성값에 따라 계속해서 색인을 한다. 그러나 ID가 6인 경우에는 더 이상 색인을 하지 않는다. 지역 문서의 가중치가 α인 링크가 지시하는 원격 문서에 포함된 링크가 가중치값을 가질 경우, 해당 링크가 지시하는 원격_원격 문서의 용어의 가중치값은 원격 문서의 링크 가중치값으로 계산되며, 이 값이 지역 문서의 용어 가중치값에 계산될 때는 지역 문서에 있는 링크의 가중치값 α을 곱하여 계산된다

3.3 색인 구조

본 논문에서 제안하는 정보검색 시스템을 위한 색인 구조는 역 화일[1]과 동등한 구조를 가진다. 우선 색인 화일에는 문서의 내용을 대표하는 색인어들을 B-Tree 형식으로 저장하고 색인어가 가리키는 포스팅 화일에는 실제 해당 색인어가 나타난 문서화일에 대한 정보를 가지고 있다. (그림 1)은 본 논문에서 제안하는 점

보검색 시스템의 색인 화일과 포스팅 화일 에 대한 구조이다.



(그림 1) 색인 화일 구조

색인 화일의 구조는 색인어, 노드수, 주소로 이루어진다. 색인어는 사용자의 질의 단어를 효율적으로 탐색하기 위하여 전체 데이터 화일에서 추출한 색인어가 들어있다. 노드 수는 해당 색인어가 포스팅 화일에 가지고 있는 노드 수를 의미하며 이는 해당 색인어가 출현한 실제 문서의 개수이다 주소는 포스팅 화일에서 해당 색인어에 대한 정보를 가지고 있는 첫 번째 노드 주소이다 포스팅 화일은 URL, 지역_빈도수, 원격_빈도수, 다음노드로 구성된다. URL은 해당 색인어가 나타난 문서 화일에 대한 URL 또는 문서 화일의 특정지점을 가리키는 XPointer가 저장된다 지역_빈도수는 색인어가 URL이 가리키는 지역 문서에서 몇 번 출현했는지를 계산하여 저장한다. 원격_빈도수는 지역 문서 안의 각 링크들이 가리키는 원격 문서에서 색인어가 몇 번 출현했는지를 계산하여 링크의 식별자(ID)별로 저장한다. 원격_빈도수를 식별자별로 저장하면, 총 횟수가 같은 문서가 여러 개 나왔을 때 각 식별자의 빈도를 이용하여 순위를 조정할 수 있다. 다음노드는 해당 색인어에 대한 포스팅 화일의 다음 노드를 가리킨다.

3.4 색인 과정 및 알고리즘

하나의 문서를 색인하기 위해 먼저, 문서에서 (지역 문서에서 모든 원격 문서까지) 나타나는 모든 링크 정보를 저장하기 위하여 임시 화일로서 링크정보 파일을 만든다. (그림 2)는 링크정보 파일에 대한 구조로서 element, id, frequency, href, remote_link, insert_link로 구성된다

하나의 문서에 대한 색인 과정에 들어가면 먼저, 문서의 처음부터 단어을 찾아 색인 파일에서 해당 단어의 등록 여부를 확인한 후, 없으면 등록하고 포스팅 파일에 해당 문서의 url과 지역_빈도를 저장한다. 계속해서 단어가 나타날 때마다 이 과정을 반복한다. 문서 안에서 링크를 만나면 링크에 대한 정보를 링크정보 화일에 저장한다. 링크를 링크정보 화일에 저장할 때에는 링크가 나타난 element를 element 필드에 저장하고 링크 식별자 테이블에서 해당 링크에 대한 식별자(ID)와 가중치를 구해서 id와 빈도 필드에 저장한다 href 필드는 해당 링크가 가리키고 있는 xlink 또는 xpointer 지점을 저장한다. 또한 해당 링크의 속성중 remote_link의 값을 remote_link 필드에 저장하고 이 값이 "NO"일 경우 insert_link 필드에 insert_link 속성에서 정의한 id들을 저장한다 계속해서 지역 문서가 끝날때까지 색인을 반복한다 지역 문서에 대한 색인이 끝나면 링크정보 화일에 있는 첫 번째 링크부터 색인을 시작한다.

링크정보 화일

element	id	frequency	href	remote_link	insert_link

(그림 2) 하나의 문서 색인시 필요한 임시 화일

링크정보 파일의 하나의 링크에 대한 색인시, 링크가 가리키는 원격 문서에서 단어을 찾아 색인화일에서 해당 단어의 등록 여부를 확인한 후, 없으면 등록하고 포스팅 파일에 해당 문서의 지역 문서에 대한 url과 링크의 가중치값을 원격_빈도에 저장한다. 원격 문서에서 또다시 링크를 만나면 이 링크는 지역 문서의 링크 속성중 REMOTE_LINK와 INSERT_LINK의 값에 따라 링크정보 화일에 등록할지를 판단한다. 링크의 가중치가 0이 아니면 링크정보 화일에 등록하고 계속해서 원격 문서의 끝까지 색인을 계속한다. 하나의 원격 문서에 대한 색인을 마치면, 다음 링크가 가리키는 원격 문서의 내용을 색인한다 링크정보 화일에 더 이상의 링크가 없을 때까지 계속한다. (그림 3)은 이에 대한 알고리즘이다

```
indexing(local_document) {
    for a document (/* 하나의 지역 문서를 색인하여
     색인화일에 저장하고 링크를 링크정보화일에 저장*/
    get a item from a document;
        if(the item is a indexing word) {
            if(the word is new)
                {add the item into data_file;}
            add_local_frequency(item);  }
        else if(the item is a link)
```

```
            insert_to_linkFile(item),
    }
    }
    for a linkFile {   /* 링크정보화일의 링크가 지시하
                          는 원격 문서를 색인한다 */
        get remote_docs from xlink or xpointer,
        for the remote_docs {
            get a item from a remote_docs:
            if(the item is a indexing word) {
                w = evaluate_weightOfLink(),
                if(the word is new)
                    {add the item into data_file.}
                add_remote_빈도(w),    }
            if(the item is a link) {
                w = evalute_weightOfRemoteLink(),
             if(w != 0) insert_to_linkFile(item);
}   } } }
```

(그림 3) 색인 알고리즘

4. 분석 및 평가

여기서는 기존의 색인 방법과 본 논문에서 제시하는 XML 링크 정보를 이용한 색인 방법을 비교 분석하기 위하여 웹상에 존재하는 검색기 중에서 'Altavista'를 이용한 검색 결과를 분석하였다. 검색기의 대상이 주로 HTML 문서이고 실험에 이용할 적당한 XML 문서 집합을 찾지 못해 HTML 문서를 이용하였다. 이를 위해 HTML 문서에서 문서의 구조는 XML 구조라고 가정하였고 문서내의 링크는 XML 링크로 정의되었다고 가정하였다. 문서의 링크에 대한 가중치를 부여하는 방법으로서 여기서는 α, β, γ, δ를 1, 0.9, .08, 0.5로 하였다. 분석을 위하여 'Altavista' 검색기에 'xml'이라는 질의어로서 검색을 하였다. 검색 결과 상위 8개의 순위에 해당되는 문서의 URL을 살펴보면 <표 5>와 같다.

〈표 5〉 'xml'에 관련된 문서들

문 서	'xml' 질의어에 대한 검색 결과
문서_1	http://xml.css.co.kr/
문서_2	http://xml.t2000.co.kr/spec_faq/spec_korean.html
문서_3	http://www.www-kr.org/
문서_4	http://indra.snu.ac.kr/~owen/XML/xml.html
문서_5	http://www.ibase.co.kr
문서_6	http://xml.t2000.co.kr/faq/index.html
문서_7	http://dblab.comeng.chungnam.ac.kr/
문서_8	http://members.jworld.net/aster/web/dhtml-xml.html

<표 5>에서의 검색 결과는 'xml'이라는 질의어에 대한 한글 내용을 검색한 결과이다. 검색기 "Altavista"의 검색 알고리즘을 정확히 알 수 없어서 검색 결과의 순위와는 무관하게 8개의 문서에 대해 기존의 일반적인 색인기법과 링크 정보를 이용한 색인 기법을 적용하였다. <표 5>의 검색 결과에는 상관없이 일반적으로 xml에 대한 좋은 정보를 가진 것으로 많이 알려진 사이트들을 살펴보면 <표 6>과 같다.

〈표 6〉 많이 알려진 XML 사이트들

	Top XML Sites
사이트_1	*IBM's XML Web Site [www.ibm.com/developer/xml]*
사이트_2	XML Developer Center[*msdn.microsoft.com/xml/*]
사이트_3	OASIS home page [www.oasis-open.org]
사이트_4	XML.org [www.xml.org]
사이트_5	The World Wide Web Consortium's XML [www.w3.org/xml/]
사이트_6	XML.COM [www.xml.com/xml/pub]
사이트_7	The XML Working Group FAQ [www.ucc.ie/xml/]

일반적으로 xml에 관련된 많은 문서에서 <표 6>의 사이트로 링크가 설정되어져 있다. <표 5>의 8개의 문서에 포함된 링크의 종류를 보면, <표 5>의 다른 문서, <표 6>의 사이트, 기타 다른 사이트로 설정되어져 있다. 질의어가 'xml'이었으므로 문서내의 링크들에 대해서도 xml에 관련된 일부의 링크들만 존재하는 것으로 가정하였고 링크의 식별자를 임의로 가정하였다. 문서내의 링크중 존재하는 것으로 가정한 링크의 URL과 링크 식별자는 <표 7>과 같다.

〈표 7〉 8개의 문서에 대한 링크 구조

문 서	링크의 URL[링크의 식별자]	remote_link 값
문서_2	사이트_3[6]	"no"
문서_3	사이트_3[6] , 문서_7[5]	"no"
문서_4	문서_4_1[6]	"yes"
문서_6	사이트_5[6], 사이트_7[6]	"no"
문서_8	사이트_2[6], 사이트_5[6] 사이트_6[6]	"no"
문서_4_1	사이트_5[6], 사이트_7[6], 문서_2[6], 문서_6[5]	"no"

<표 7>의 링크 구조를 가진 8개의 문서에 대해 기존의 색인 기법 즉, 링크는 무시하고 모든 문서를 독립적인 문서로 간주하여 색인한 기법과 본 논문에서 제시한 링크 정보를 이용한 색인 기법 즉, 문서내의 링크가 가리키는 원격 문서에서도 색인을 하는 기법을

적용하여 문서내 색인어 발생 빈도수를 조사하였다. 분석 결과, 'xml'이라는 색인어에 대해 각 문서에서의 발생빈도 순위는 <표 8>과 같다.

〈표 8〉 'xml' 색인어의 발생 빈도 순위

문 서	발생빈도 순위	
	기존의 색인	링크기반 색인
문서_1	⑦	⑦
문서_2	①	⑤
문서_3	⑥	④
문서_4	⑤	①
문서_5	⑧	⑧
문서_6	②	③
문서_7	④	⑥
문서_8	③	②

결과적으로, 하나의 문서에서 기존의 색인 기법과 링크기반 색인 기법에 의한 색인 결과는 색인어 발생 빈도수가 서로 다르게 나타난다는 것을 알 수 있다. 따라서, 이러한 색인어 발생 빈도수를 문서의 순위 매김에 이용하면, 문서의 순위에 영향을 줄 수 있다. 또한 분석에 이용한 예제가 HTML 문서인 관계로 실제 XML 링크의 모든 식별자를 이용하지 못하였다. 실제 XML 링크 ID(1), ID(2), ID(3)이 많이 존재하는 XML 문서의 경우에는 실제 문서를 구성하고 있는 내용과 그 문서가 웹 브라우저상에 디스플레이되는 내용이 많이 다를 수 있다. 따라서, 기존의 색인 방법에서는 실제 문서를 구성하고 있는 내용을 기반으로 색인을 하기 때문에 나중에 그 문서가 웹 브라우저상에 디스플레이될 때는 색인어 발생 빈도 순위와 내용이 많이 다를 수 있다. 본 논문에서 제시하는 링크 정보를 이용한 색인 기법은 문서가 웹 브라우저상에 디스플레이되는 시점을 기반으로 색인을 하기 때문에 색인어에 대한 발생 빈도 순위를 매길 때, 보다 근접한 순위 매김을 할 수 있다.

5. 결론 및 향후과제

본 논문에서는 XML의 링크 정보를 이용한 새로운 정보검색 시스템을 위한 색인 기법을 제시하였다. XML 링크에 지역 문서에서 원격 문서에 있는 링크를 제어할 수 있는 새로운 속성을 정의하여 각 속성별로 정의된 서로 다른 링크들에 대해 고유한 식별자를 부여하고 각 식별자별로 링크의 가중치를 부여하였다. 링크의 가중치는 해당 링크가 지역 문서에 있을때와 원격 문서에 있을 때 그 값이 달라진다. 따라서 하나의 문서를 색인할 때, 링크된 원격 문서는 링크의 가중치 값을 기준으로 색인을 하며, 원격 문서 안의 링크들에 대해서는 그 속성에 따라 계속해서 색인할 지의 여부를 판단하여 해당하는 가중치 값으로 계속해서 색인을 한다. 이 방법을 XML 문서에서 기존의 방법과 비교 분석해 본 결과 문서에서의 링크에 대한 구조 정보를 무시하고 각 문서를 독립적으로 간주하여 색인한 기존의 방법에 비해서 본 논문에서 제시하는 링크 정보를 이용한 색인 방법은 XML 문서의 종류에 따라 사용자의 질의에 보다 근접한 문서를 색인할 수가 있었다.

향후 연구과제로는 대규모 XML 문서 컬렉션을 대상으로 링크 기반 검색의 효율성을 검증하고 아울러 문서의 논리적 구조 정보를 이용한 문서의 부분 검색 기능과의 통합을 통해서 XML 문서 검색기를 개발하는 것이다.

참 고 문 헌

[1] William B Frakes and Ricardo Baeza-Yates, "Information Retrieval Data Structures & Algorithms," Prentices Hall, 1992.

[2] W.Bruce Croft and Howard Turtle, "A retrival model for incorporating hypertext links," Hypertext'89 proceedings, pp.213-224, 1989.

[3] Dario Lucarella "A model for hypertext-based information retrival," In Hypertext Concepts systems, and Applications, Eds.Rizk, Streitz, and Andrie, 1990.

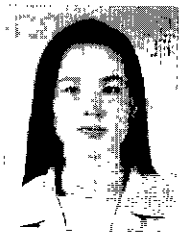
[4] Jacques Savoy, "An extended Vector-processing scheme for searching information in hypertext systems." Information processing&Management, Vol. 32, No 2, pp.155-170, 1996.

[5] 김동욱, 류준형, 주원균, 맹성현, "링크 정보를 이용한 검색 신뢰도의 향상", 한국정보과학회 춘계 학술발표논문집, Vol.25, No.1, pp.446-448, 1998.

[6] 맹성현, 주종철, "문서구조화와 정보검색", 한국정보과학회지, 제16권 제8호, pp.6-15. 1998.

[7] Rohit Khare, Adam Rifkin "XML : A door to Au-

tomated Web Applications," IEEE Internet Computing, pp.78-87, July & August 1997.
[8] Ronald96, Ronald C. Tumothy A. Douglass. Audrey J. Turner "Readme.1st SGML for Writers and Editors," PH, 1996.
[9] D. Connolly and J. Bosak, "Extensible Markup Language(XML)," 1997, http //www.w3c.org/XML/
[10] 정희경, "차세대 웹 문서 표준 XML", 한국정보처리학회지, 제6권 제3호, pp.25-35,1999.
[11] 나홍석, 최진석, 김창화, 백두권, "차세대 웹 상에서의 문서 교환 및 검색을 위한 프레임 워크", 한국정보처리학회지, 제6권 제3호, pp.52-61, 1999.
[12] W3C Working Draft 21-February-2000, "XML Linking Language(XLink)," http : //www.w3 org/TR/xlink
[13] W3C Working Draft 21-February-2000, "XML Pointer Language(XPointer)," http . //www.w3.org/TR/WD-xptr
[14] 신봉기, 김영환, "인터넷 정보검색 서비스 동향", 한국정보과학회지, 제16권 제8호, pp 16-20, 1999.
[15] Light, R "Presenting XML," Sams.net Publishing. 1997.
[16] "SGML/XML '97 Conference Proceedings," pp.8-11, December, 1997, Sheraton Washington Hotel, Washington, D.C.



김 은 정

e-mail : ejkim@base.gsnu ac.kr
1989~1993년 LG전자 영상미디어 연구소 연구원
1996년 경상대학교 전자계산학과 졸업(석사)
1998년 경상대학교 전자계산학과 박사수료
관심분야 : 웹 프로그래밍 언어, 디지털 라이브러리, 정보검색

배 종 민

e-mail : jmbac@base.gsnu.ac.kr
1980년 서울대 사범대학 수학과 졸업(학사)
1983년 서울대 대학원 계산통계학과 졸업(석사)
1995년 서울대 대학원 계산통계학과 졸업(박사)
1982년~1984년 한국 전자통신연구소 연구원
1984년~현재 경상대학교 컴퓨터과학과 교수
관심분야 . 웹 프로그래밍 언어, 디지털 라이브러리, 정보검색