사용자 질의를 이용한 XML 태그의 가중치 결정

우 선 미[†] · 유 춘 식[††] · 김 용 성[†††]

요 약

보다 효과적인 색인어 추출 및 색인어 가중치 결정을 위하여 문서의 내용뿐만 아니라 구조를 이용하여 색인을 추출하는 연구가 이루어지고 있다. 이러한 연구들 대부분이 XML 태그의 중요도가 아닌, 문맥상의 단락에 대한 중요도를 계산하거나 HTML 문서 태그의 중요도를 결정하는 연구들이다. 이러한 기존 연구들은 대부분이 객관적인 실험을 통해서 중요도를 입증하기보다는 상식적인 관점에서 단순한 수치로 중요도를 결정하고 있다. 본 논문에서는 웹 문서 관리를 위한 표준으로 자리잡아가고 있는 XML 문서의 태그 정보를 이용한 자동색인을 위하여, 논문을 구성하는 주요 태그의 가중치를 계산하는 방법을 제안한다. 보다 객관적인 가중치 결정을 위하여 사용자의 질의에 바탕을 둔 사용자의 검색 행위를 반영한다. 그리고 기존 방법을 적용하여 계산된 색인어 가중치를 이용한 검색성능과 비교함으로써 본 논문에서 제안한 방법을 적용하여 계산된 색인어 가중치의 효과를 검증한다.

키워드 : XML 태그 가중치, 자동 인덱싱, 정보검색

Weighting of XML Tag using User's Query

Seon-Mi Woo[†] · Chun-Sik Yoo[††] · Yong-Sung Kim[†††]

ABSTRACT

XML is the standard that can manage systematically WWW documents and increase retrieval efficiency. Because XML documents have the information of contents and that of structure in single document, users can get more suitable retrieval result by retrieving the information of content as well as that of logical structure. In this paper, we will propose a method to calculate the weights of XML tags so that the information of XML tag is used to index decision. A proposed method creates term vector and weight vector for XML tags, and calculates weight of tag by reflecting user's retrieval behavior (user's query). And it decides the weights of index terms of XML document by reflecting the weights of tags. And we will perform an evaluation of proposed method by comparison with existing researches using weights of paragraphs.

Key Words : Weight of XML Tag, Automatic Indexing, Information Retrieval

1. 서 론

하이퍼 텍스트 개념에 바탕을 둔 웹이 1990년대 중반부터 폭발적으로 성장함에따라 인터넷은 빠른 속도로 정보 통신 구조의 표준으로 자리를 잡게 되었다. 웹의 발달과 인터넷의 보편화로 인하여 정보의 양이 기하급수적으로 증가함에 따라, 자신이 원하는 정보를 얻기가 점점 어려워졌다. 따라서 보다 효과적으로 웹 문서에 대한 색인을 추출하고 검색 편의성을 제공하는 연구가 필요하다. 이러한 문제점을 해결하기 위한 대안 중의 하나가 정보를 XML(eXtensible Markup Language) 형태로 관리하는 것이다. XML은 문서의 구조정보를 제공할 뿐만 아니라, XML 태그(tag)는 데이터를 해석하는 데에 사용할 수 있기 때문에 XML의 역할과 중요성이 증대되고 있다. HTML이 하나의 고정된 DTD (Document Type Definition)를 사용하는 것과는 달리 XML은 논리적 구조를 나타내는 여러 DTD를 사용할 수 있다. XML 문서는 하나의 문서에 내용 정보와 구조 정보를 가지고 있기 때문에 기존의 내용 정보에 대한 검색뿐만 아니라 논리적인 구조 정보를 이용하여 검색할 수 있는 기능도 필요하다[3, 15, 20]. 또한 문서의 내용 정보뿐만 아니라 문서의 구조 정보를 이용하여 색인을 추출하고 색인어 가중치를 계산할 수 있다면 검색 효율성을 높일 수 있을 것이다. 태그의 중요도에 관한 연구가 일부에서 이루어지고 있지만, 대부분의 연구들이 XML 태그의 중요도가 아닌, 문맥상의 단락에 대한 중요도를 계산하거나[1, 6, 13, 14], HTML 문서 태그의 중요도에 관한 연구[4]이다. 또한 이들 연구들은 객관적인 실험을 통해서 중요도를 입증하기보다는 상식적인 관점이나 전문가의 휴리스틱(heuristic)에 의하여 단순한 수

† 정 회 원 : 전북대학교 전북지역전자정보사업단 기금교수
†† 정 회 원 : 전북대학교 전산통계학과 이학박사
††† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2005년 2월 1일, 심사완료 : 2005년 4월 14일

치로 중요도를 결정하고 있다. 따라서 보다 효율적으로 XML 문서의 자동색인을 수행하기 위하여 본 논문에서는 사용자의 검색 행위 관찰을 통한 XML 태그의 가중치 결정 방법을 제안한다.

XML 문서의 자동색인 및 색인어 가중치 결정을 위한 태그의 가중치를 계산하기 위하여, 본 논문에서는 텍스트 XML 문서 집단(논문과 연구보고서)을 대상으로 사용자의 검색 행위를 알아본다. XML 문서에서 색인어를 추출하여 주요 태그마다 태그 용어 벡터와 태그 가중치 벡터를 생성한다. 이때 사용하는 태그는 문서를 작성할 때 자주 사용하는 제목, 목차, 저자, 키워드, 요약, 서론, 본론, 결론, 참고문헌이다. 사용자가 입력한 질의를 질의 벡터로 구성하여, 이를 태그별로 구성된 태그 용어 벡터와 비교하여 태그마다 생성한 태그 가중치 벡터에 반영한다. 다음으로 사용자의 검색 행위를 반영한 태그별 태그 용어 벡터와 태그 가중치 벡터를 이용하여 XML 문서의 태그 가중치를 계산한다. 실험 대상 문서는 웹에서 검색한 컴퓨터과학 및 정보통신 분야의 논문과 연구보고서를 이용한다. XML로 표기되지 않는 것들은 XML로 재구성하여 사용한다. 실험 평가는 두 가지 측면에서 실시한다. 첫 번째 실험은 본 논문에서 제안하는 방법으로서 태그의 가중치를 결정하는 실험이고, 두 번째 실험은 계산된 태그의 가중치를 반영하여 결정한 색인어 가중치의 성능 실험이다. 색인어 가중치의 성능비교는 (1) 일반 $TF \cdot IDF$를 이용하여 색인어 가중치를 결정하는 방법, (2) 논문의 단락 즉, 제목, 초록, 키워드, 서론, 관련연구, 내용, 실험, 결론, 감사의 글, 참고문헌에서 색인어 빈도수를 이용하여 단락의 가중치를 계산하고 있는 방법[4], 그리고 (3) 본 논문에서 제안하는 방법을 대상으로 수행한다.

먼저 2장에서 XML 문서의 구조 및 태그의 종류를 간단히 살펴보고, 문서의 단락 정보를 정보검색에 이용하기 위한 기존 연구를 소개한다. 3장에서는 본 논문에서 제안하는 XML 태그의 가중치 계산 방법을 설명한다. 4장에서는 본 논문에서 제안하는 방법으로 계산된 XML 태그 가중치가 검색성능에 미치는 영향을 알아보기 위한 실험을 수행하고, 마지막으로 5장에서 결론과 함께 향후 연구 방향에 관해 기술한다.

2. 관련 연구

본 장에서는 XML 문서의 구조를 간단히 알아보고, 기본적인 색인어 가중치 결정 방법과 문서의 일부에 중요도를 주기 위한 기존 연구들을 살펴본다.

2.1 XML 문서의 구조

XML(eXtensible Markup Language)은 구조화된 문서를 표현하고 상호 교환하기 위한 전자문서 표현형식이다. 또한 XML은 문서의 개념적인 논리 구조와 내용 구조를 기술할 수 있으며, 복잡한 구조와 데이터를 포함하는 멀티미디어 문서와 하이퍼링크를 포함하는 하이퍼미디어 문서까지도 작성할 수 있다[19]. 또한 XML 문서는 ASCII 텍스트형의 태그를 사용하기 때문에 시스템이나 응용 프로그램에 상관없이 문서의 상호 교환이 가능하다는 장점을 가진다[9]. XML은 문서에서 사용할 수 있는 태그와 속성 세트를 문서형 정의(Document Type Definition; DTD)를 통해 정의한다. XML은 실제의 문서 정보를 포함하고 있는 엘리먼트(element), 엘리먼트에 포함되어 추가적인 정보를 제공하는 속성(attribute), 약어나 이진 데이터를 사용하기 위한 엔티티(Entity), 응용 프로그램을 활용하기 위한 처리명령(processing instruction), XML 프로세서가 해석하지 않는 설명문인 주석(comment), 그리고 문자열을 일반 텍스트로 인식하도록 하는 CDATA 섹션(CDATA Section)으로 구성된다. XML 문서의 구조정보는 부모 엘리먼트와 자식 엘리먼트간의 계층정보, 동일한 부모 엘리먼트를 갖는 형제 엘리먼트들에 대한 연결자와 반복연산자에 의해 표현된다. 그러므로 기준 엘리먼트로부터 특정 엘리먼트에 대한 계층정보와 순서 정보를 손쉽게 구할 수 있다[2, 5, 12, 18, 20]. 이렇게 구한 순서정보는 각 엘리먼트에 유일하게 할당된 값이기 때문에 직접 특정 엘리먼트에 접근할 수 있다. 또한 DTD에 나타난 반복연산자에 의한 반복적인 엘리먼트의 사용이 가능하여 엘리먼트간의 순서정보를 유지함으로써 특정 엘리먼트를 검색하는데 유용하게 이용된다.

2.2 색인어 가중치 결정

색인어에 가중치를 부여하는 목적은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라 색인어로서의 상대적 가치를 표현하기 위함이다. 자동색인 기법에서 색인어 가중치 결정은 주로 통계적 기법을 이용하는데, 통계적 기법의 통계적 기준은 대부분 색인어의 출현빈도에 근거하고 있다. 색인어의 출현빈도에는 용어빈도(TF : Term Frequency), 문서빈도(DF : Document Frequency), 장서빈도(CF : Collection Frequency)로 구분할 수 있다. 용어빈도는 색인 대상이 되는 각 문서 i에 특정한 용어 k가 출현한 횟수를 나타낸다($TF = f_{ik}$). 문서빈도는 특정한 용어 k가 출현한 문서 수 $DF = \sum_{i=1}^{l} b_{ik}$이며, $f_{ik} \geq 1$일 때 $b_{ik} = 1$, $f_{ik} = 0$일 때 $b_{ik} = 0$이다. 장서빈도는 특정 용어 k가 전체 문서집단 내에 출현한 총 빈도로서 $CF = \sum_{i=1}^{l} f_{ik}$ 가 된다. 용어빈도를 문서빈도나 장서빈도로 나누어 줌으로써 빈도값을 표준화시킬 수 있다[1, 10, 11]. 역문헌빈도란 용어빈도를 문서빈도로 나누어주는 것을 말한다. 대표적인 공식으로서 스파크 존스(Spark Jones)가 제시한 역문헌빈도의 공식은 (식 1)과 같다[10].

$$W_{ik} = log_2 \frac{n}{DF} + 1 = log_2(l) - log_2(DF) + 1 \qquad (1)$$

단, W_{ik} : 문서 i에서 용어 k가 갖는 색인어로서의 가중치
l : 전체 문서

스파크 존스가 수행한 실험을 살펴보면 어떤 용어가 문서 집단에 많이 나타나면서 해당 용어가 나타나는 문서의 수가 적을수록 색인어로서의 가치가 크다는 것을 알 수 있다[9]. 이러한 방법은 문서 내에서의 특정 용어의 중요도는 해당 문서 내의 출현 빈도와 비례하고 총 출현 문서의 개수와는 반비례하는 특성을 활용하여 중요 용어를 추출할 수 있다. 그러나, 용어의 위치 정보나 문장 사이의 구분과 같은 구조적 정보를 고려하지 못한다는 단점이 있다[11].

2.3 문서의 위치 중요도에 관한 연구

색인어의 가중치를 계산할 때 색인어의 가중치가 부여된 문서의 단락 위치 즉, 제목, 초록, 키워드, 서론, 관련연구 등에서의 용어 빈도를 이용한 가중치를 계산하는 방법[4]과 HTML 태그에 가중치를 부여하여 색인어 가중치를 결정하는 방법[3]에 대해 간단히 살펴본다.

2.3.1 문서 위치를 고려한 가중치 계산

[4]에서는 문서의 제목이나 키워드 부분에서 추출된 용어는 서론이나 관련연구 부분에서 추출된 용어보다 더 중요도를 가지고 있다는 가설에 근거하여 문서를 구성하는 단락의 가중치를 계산하고 있다. 문서의 각 단락에서 추출된 용어 t_k는 (식 2)와 같이 용어빈도와 문서 위치에 부여된 가중치를 이용하여 계산한다.

$$Sup_{t_k} = \frac{sup_{t_k}}{MAX\{sup'_{t_k}\}} \quad (2)$$

단, $Sup'_{t_k} = \sum tf_{ik} \cdot W_{ik}$ 에 의하여 계산

tf_{ik} : 문서i 의 위치 t에 있는 용어 k의 빈도

W_{ik} : 문서i 의 용어 k에 대한 위치 가중치

<표 1>은 문서의 위치에 따른 가중치를 나타내고 있는데, 이러한 결과를 얻기까지의 방법은 자세히 서술되어 있지 않다. 연구자의 경험지식에 의한 결과가 아닌 객관적인 근거를 바탕으로 한 중요도 결정을 위한 연구가 필요하다.

2.3.2 태그 정보를 이용하여 가중치를 부여하는 방법

웹 문서가 태그로 이루어진다는 특징에 착안하여 색인어로서 중요한 내용을 담고 있는 HTML 태그와 덜 중요한 내용을 담고 있는 HTML 태그를 차별하여 가중치를 부여하고

〈표 1〉 문서의 위치에 따른 가중치

분 류	가중치	분 류	가중치
제목	1.9	내용	1.5
초록	1.8	실험	1.2
키워드	2.0	결론	1.2
서론	1.3	감사의 글	1.0
관련연구	1.6	참고문헌	1.4

〈표 2〉 HTML 태그의 가중치가 높은 순위

순위	태그	순위	태그
1	<a>	12	<big>
2		13	
3	<h1>,<h2>	14	<u>
4	<h3>,<h4>	15	<strike>
5	<h5>,<h6>	16	<cite>
6	font size 6이상	17	<i>
7	font size 4, 5	18	<var>
8	font size 1~3	19	
9	<blink>	20	<tt>
10	<marquee>	21	
11	<b>	22	

〈표 3〉 가중치 테이블 기준

분류	가중치	태 그 내 용
최고어	10	<a> +
링크	9	<a> + 을 제외한 태그
제목효과1	8	+ <a>를 제외한 태그
제목효과2	7	<h1>, <h2>, font size 6이상 + blink, marquee
제목	6	<h1>, <h2>, font size 6이상 + 그 이하 point 태그
강조효과	5	태그 point 2,3 글 크기 + blink, marquee
효과	4	blink, marquee
강조1	3	태그 point 3 + 태그 point 2
강조2	2	태그 point 2 + 태그 point 1
본문	1	본문

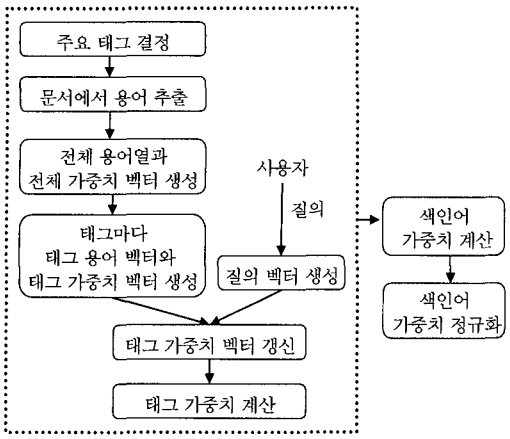
색인어 가중치를 결정하는 방법이 있다[3, 6, 17]. [3]은 <표 2>와 같은 순위를 고려하여 1~9의 범위를 갖는 가중치 값을 부여하였는데, font size가 1~3인 경우와 <h5> 및 <h6>은 기준 가중치로서 1을 주고, 링크가 된 곳은 최고의 가중치인 9를 부여하고 있다.

다음으로 제목, 주제, 효과, 강조로 나누어 2~8까지의 가중치를 주었다. 태그가 2개 이상이 중복될 경우에는 <표 3>의 가중치 테이블을 기준으로 하여 순위가 높은 태그에만 가중치를 부여하고 있다.

이 연구는 하이퍼텍스트의 특성만을 고려하기 때문에 링크 정보에 상당히 의존하여 태그의 중요도를 결정하고 있다. 문서의 제목일 경우엔 주제를 표현한다고 볼 수 있지만, 폰트 사이즈로서 제목을 구분하므로 일반적인 문서의 소제목 또한 높은 가중치를 갖게 된다. 인문 서적의 경우와는 달리 논문이나 연구보고서와 같은 문서의 경우에는 소제목이 주제를 제대로 표현한다고는 볼 수 없다. 예를 들어 논문의 소제목의 경우, 관련 연구, 서론, 본론, 시스템 설계, 구현 환경과 같이 주제와 관련이 적은 경우에도 높은 가중치를 갖게 된다. 따라서 이 방법은 논문 구성이나 연구보고서에 대한 자동색인에는 그다지 적합하지 않다.

3. XML 태그 가중치를 이용한 색인어 가중치 계산

본 장에서는 논문이나 연구보고서와 같은 종류의 문서들



(그림 1) XML 태그의 가중치 결정 과정

을 대상으로 XML 문서를 구성하는 태그의 가중치를 계산하고, 이 태그의 중요도를 이용하여 색인어 가중치를 계산하는 방법을 기술한다. XML 태그의 가중치를 결정하는 과정은 (그림 1)의 박스 안과 같다.

먼저 논문을 구성하는 XML 문서의 여러 태그들 중에서 가중치 계산에 이용할 태그를 결정하기 위하여 설문조사를 실시하였다. 대학원생 중에서 논문 표현을 위한 XML 태그를 알고 있는 30명을 대상으로 "XML 문서로 된 논문을 검색할 때 태그별로 검색이 가능하다면 어떤 태그로 검색하겠는가?"라는 질문을 하였다. 설문 결과 저자, 출판년도, 출처, 제목, 목차, 초록, 키워드, 서론, 본론, 결론, 참고문헌이라는 태그를 얻을 수 있었고, 주제 색인과 비주제 색인으로 구별한 결과는 <표 4>와 같다.

<표 4> 설문조사로 선정한 XML(논문) 태그

비주제 색인		주제 색인	
No.	tag	No.	tag
1	저자	1	제목
2	출판년도	2	목차
3	출처	3	초록
		4	키워드
		5	서론
		6	본론
		7	결론
		8	참고문헌

주제 색인은 정보자료의 주제를 나타내는 요소를 색인으로 선택하는 색인을 말하며, 비주제 색인은 저자명, 기관명, 출판년, 프로젝트명, 보고서 번호 등과 같이 주제와는 직접적으로 관계없는 요소를 색인어로 선택하는 색인이다[11]. 본 논문에서는 본문의 내용을 대상으로 하여 색인어를 자동으로 추출하고 가중치를 결정하는 것이 목적이므로, 주제 색인어를 추출할 수 있는 제목, 목차, 초록, 키워드, 서론, 본론, 결론, 참고문헌 태그를 선택하여 각 태그의 가중치를 결정한다. 그리고 본론은 관련연구, 실험평가, 구현 및 설계 등의 내용을 포함한다.

3.1 전체 가중치 벡터와 태그 벡터 생성

태그 가중치를 계산하기 위하여 테스트 집합을 이용하여 사용자들의 검색 행위를 관찰하였다. 사용자가 입력한 질의어와 일치하는 색인어가 위치하는 태그를 조사한 결과, 요약, 서론, 본론, 결론 태그에 질의어와 일치하는 색인어가 가장 많이 포함되었다. 그러나 이러한 정보만으로 태그의 가중치를 결정하기엔 미흡한 점이 많다. 즉 본문의 경우엔 추출된 용어 자체가 많아서 당연히 질의와 일치되는 경우가 많다. 본 논문에서는 질의와 태그 내의 용어가 일치하는 빈도수와 원래 용어의 가중치를 기반으로 하여 태그의 가중치를 결정한다. 자세한 방법은 다음 소절에서 설명한다.

3.1.1 전체 용어열과 전체 용어 가중치 벡터 생성

태그의 가중치 계산과 색인어 선정에 사용하기 위하여 문서집합에서 추출한 용어를 이용하여 전체 용어열 $T_Total = (dt_1, dt_2, ..., dt_n)$와 전체 용어 가중치 벡터 $W_Total = (dw_1, dw_2, ..., dw_n)$을 생성한다. 이때 n은 문서집합을 구성하는 문서에서 추출한 전체 용어의 개수이다. 전체 용어 가중치 벡터는 전체 용어열과 쌍을 이루는 벡터로서 각 용어에 대한 가중치를 나타낸다. 전체 용어 가중치 벡터의 값은 역문헌빈도($TF \cdot IDF$) 방법을 이용하여 구한다.

3.1.2 태그마다 용어 벡터와 가중치 벡터 생성

선정된 태그마다 용어 벡터 $T_tag_j = (tag_t_{j1}, tag_t_{j2}, ..., tag_t_{jn})$과 가중치 벡터 $W_tag_j = (tw_{j1}, tw_{j2}, ..., tw_{jn})$을 생성한다. 이때 j는 <표 4>에서 주제 색인의 NO(번호)를 나타낸다. 즉, 제목태그 용어 벡터와 제목태그 가중치 벡터는 각각 T_tag_1와 W_tag_1이 된다. 태그 용어 벡터와 태그 가중치 벡터의 크기는 전체 용어 가중치 벡터와 같다. 태그 용어 벡터의 값은 해당 태그에서 추출된 용어의 경우엔 1값을, 그렇지 않은 경우엔 0값을 갖는다. 태그 가중치 벡터의 값은 (식 3)에 의해 계산한다. 즉 태그 가중치 벡터 W_tag_j는 T_tag_j에 대응되는 가중치 벡터로서, T_tag_j와 연산하여 해당 태그에서 출현하는 용어에 대한 가중치만을 계산하여 갖게 되고, 나머지는 0값이 된다.

$$W\_tag_{jk} = W\_total_k \cdot T\_tag_{jk} \cdot \frac{1}{DF_j} \quad (3)$$

단, W_total_k : 전체 용어 가중치 벡터의 k번째 값

T_tag_{jk} : j번째 태그의 가중치 벡터의 k번째 값

DF_j : j번째 태그 위치에서 출현한 용어의 문서 빈도

j : 1, 2, ..., 8 (1번은 제목 태그, 2번은 목차 태그,..., 8번은 참고문헌 태그)

(식 3)에서 W_tag_{jk}는 j번째 태그의 k번째 용어 가중치

를 의미하는 것으로, 여러 문서에서 출현하는 용어의 가중치를 낮춰주기 위하여 W_total_{jk}의 가중치에 T_tag_j가 출현하는 문서의 개수(문서빈도)의 역을 곱해주었다.

3.1.3 질의 가중치 벡터 생성

3.1.2절에서 태그마다 생성한 태그 가중치 벡터의 가중치에 사용자 검색행위를 반영하기 위하여 사용자가 입력한 질의를 이용하여 질의 가중치 벡터 $W_qury_m = (qw_{m1}, qw_{m2}, qw_{mn})$를 생성한다. m은 질의의 총 개수를 의미한다. 질의 가중치 벡터 또한 생성 초기에는 전체 용어 가중치 벡터와 동일하다. 질의 벡터는 20명의 사용자로 하여금 30회 이상 테스트 집합에서 검색을 수행하도록 하여 구하였다. 질의 가중치 벡터의 값은 전체 용어열 중에서 질의를 구성하는 용어와 일치하는 부분에 대응되는 값은 1이 되고, 나머지는 모두 0이 된다. 본 논문에서는 사용자가 질의로 사용하는 용어를 이용하여 사용자의 검색 행위를 관찰하여 태그 가중치에 반영하는 것이 목적이므로, 사용자가 주는 질의의 가중치를 모두 1로 하였다. 질의를 구성하는 용어에 중요도를 각기 달리하여 태그 가중치 벡터에 반영했을 경우의 검색효율에 관하여 계속 연구할 계획이다.

3.2 태그 가중치 벡터의 갱신

태그마다 태그 가중치 벡터를 생성하고, 질의 가중치 벡터가 생성되면, 태그 가중치 벡터와 질의 가중치 벡터를 비교하여 태그 가중치 벡터의 값을 갱신한다. 질의를 구성하는 용어와 일치하는 용어가 포함되어 있는 태그를 만나면 그 용어에 해당하는 태그 가중치 벡터의 가중치를 갱신한다. 가중치 갱신은 (식 4)에 의해 수행된다.

$$W\_tag_{jk} = W\_tag_{jk} + (W\_tag_{jk} \cdot W\_query_{mk}) \quad (4)$$

단, W_tag_{jk} : j번째 태그의 k번째 용어의 가중치

W_query_{mk} : m번째 질의에 포함된 k번째 용어의 가중치

질의 가중치 벡터의 값인 W_query_{mk}는 0 아니면 1이 사용되므로 질의 가중치 W_query_{mk}와 태그 가중치 벡터의 값인 W_tag_{jk}를 곱해주면 질의와 일치하는 태그의 가중치는 갱신되고 질의와 일치하지 않는 태그의 가중치는 갱신되지 않는다.

3.3 태그 가중치 계산

태그마다 생성된 태그 가중치 벡터를 갱신한 후, (식 5)에 의해 각각의 태그 가중치 V_tag_j를 계산한다. 예를 들어 제목 태그의 가중치는 V_tag_1이고, 서론 태그의 가중치는 V_tag_5가 된다.

$$V\_tag_j = \left(\sum_{k=1}^{n} W\_tag_{jk}\right) / C_j \quad (5)$$

단, n : 문서에서 추출한 용어의 총 개수, 벡터의 크기

C_j : W_tag_j에서 가중치가 0보다 큰 용어의 개수, 즉, T_tag_j에서 1인 값을 갖는 용어의 개수

$$= \sum_{k=1}^{n} T\_tag_{jk}$$

(식 5)는 사용자의 질의를 반영하여 갱신된 태그 가중치 벡터의 값을 모두 합산하는 식이다. 합산한 값을 해당 태그에 포함된 용어의 개수로 나누어 준다. 태그에 포함된 용어의 개수는 태그용어 벡터의 값이 1인 것들에 해당되므로, 태그 용어 벡터의 합산으로 나누어주면 된다. 이렇게 함으로써 태그에 포함된 용어들의 중요도를 알 수 있다.

본 논문에서는 사용자의 검색 행위를 관찰하여 태그의 가중치 계산에 반영함으로써, 저자의 주관적 경험에 의한 방법이 아닌 보다 객관적인 방법으로 XML 태그의 중요도를 계산할 수 있다.

3.4 색인어 가중치 계산

태그 가중치의 크기에 따라 태그의 중요 순위가 결정되면 태그 가중치를 이용하여 용어들의 가중치를 결정한 후, 일정 가중치 이상의 용어들을 색인어로 선정한다. 이를 위하여, 먼저 각 문서마다 용어 벡터 $T_i = (t_{i1}, t_{i2}, ..., t_{in})$와 각 문서의 용어에 대한 용어의 가중치를 나타내는 가중치 벡터 $W_i = (w_{i1}, w_{i2}, ..., w_{in})$을 생성한다. 벡터의 크기는 전체 문서 가중치 벡터의 크기와 같다. 문서 가중치 벡터의 초기값은 $TF \cdot IDF$ 방법으로 구하는데, 해당 문서에서 출현하지 않은 용어에 대한 가중치는 0이다. 다음으로 문서마다 출현하는 용어를 8개의 태그 용어 벡터와 비교하여, 일치하는 용어가 발생하면 태그 가중치인 V_tag_j를 그 용어의 가중치에 반영해 준다. 태그 가중치를 반영하여 최종 용어의 가중치를 계산하는 공식은 (식 6)과 같다.

$$W_{ik} = \sum_{j=1}^{8} | W_{ik}(1 + V\_tag_j) | \quad (6)$$

단, W_{ik} : i번째 문서의 k번째의 용어의 가중치

V_tag_j : W_{ik}이 용어가 포함되어 있는 j번째 태그 가중치

이때 용어가 여러 태그에 중복 위치한 경우, 각 태그의 중요도 값인 V_tag_j는 모두 더해지므로 색인어 가중치가 상대적으로 높아진다.

$TF \cdot IDF$값으로 정해진 초기 색인어 가중치는 [0, 1] 범위의 값이 나올 수도 있지만, 문서에서 추출된 어떤 용어가 요약, 서론, 본론에도 포함되어 있을 수 있으므로 문서의 용어가 포함된 태그 가중치를 반영하여 계산하면 최종 용어의 가중치는 [0, 1] 범위를 벗어날 수도 있다.

$$N\_W_{ik} = \frac{W_{ik} - W_{\min}}{W_{\max} - W_{\min}} \quad (7)$$

단, $W_{\min}$: W_{ik}중에서 가장 작은 값,
W_{max} : W_{ik}중에서 가장 큰 값

(식 7)에 의해 용어의 가중치를 정규화 시켜 [0, 1] 범위의 값을 갖는 각 문서에 대한 최종 색인어 가중치 N_W_{ik}를 계산한다.

4. 실험 및 평가

본 장에서는 본 논문에서 제안하는 방법에 의하여 태그의 가중치를 결정하기 위한 실험과 결정된 태그 가중치를 이용한 검색 성능을 평가하기 위한 실험 결과를 기술한다. 검색 성능은 결정된 태그 가중치를 반영하여 색인어 가중치를 결정한 후, 일반적인 검색성능 평가척도를 이용하여 성능을 평가한다. 또한 본 논문에서 제안하는 태그의 가중치를 반영하여 계산된 색인어 가중치의 성능을 확인하기 위하여 문서순위결정을 수행한 후, 상위 문서의 적합성 정도를 평가한다. 본 논문에서는 실험 평가를 위하여 정확률(precision ratio)과 재현율(recall ratio)[10] 그리고 적합률(relevance ratio)[7]을 이용하였다.

4.1 실험 환경

태그 가중치 결정과 색인어 가중치 결정 및 검색 성능을 평가하기 위한 실험 환경은 다음과 같다.

▸실험 데이터

웹에서 검색한 컴퓨터과학 및 정보통신 분야의 논문과 연구보고서를 XML로 재구성한 문서 약 300개

▸실험 참가자

컴퓨터과학 및 정보통신 분야의 석사학위 과정 이상인 전공자 33명

▸색인어 추출범위와 방법

－색인어 추출 범위 : 문서의 전체

－색인어 추출 방법 : 자동 색인 방법[8] 이용

웹에서 데이터를 얻기 위하여 일반 웹 검색엔진과 도서관 원문 서비스를 이용하였다.

4.2 실험 평가

두 가지 측면에서 실험평가를 실시한다. 첫 번째 실험은 태그의 가중치 결정을 위한 실험이고, 두 번째 실험은 태그 가중치를 반영한 검색 성능 평가를 위한 실험이다.

4.2.1 태그 중요도 결정 실험

XML로 변환된 컴퓨터과학 및 정보통신 분야의 문서에서 5개의 관심분야 별로 각각 20명씩 전공자로 하여금 평균 15회 이상 질의를 입력하여 검색을 실시하게 하였다. 이때 한 사람이 여러 관심분야를 가질 수 있도록 하였다. 총 참여 인원은 33명이고, 33명이 평균 3개의 관심분야를 테스트하게 하였으며 총 질의 횟수는 약 900번이다. 그리고 본 논문에서 제안하는 과정에 의해 태그 가중치를 계산한 후, 각 분야별로 평균값을 구하면 <표 5>와 같다.

〈표 5〉 태그 가중치 실험 결과

No.	tag	중요 순위	태그가중치(%)
1	제목	2	23%
2	목차	8	4%
3	초록	3	18%
4	키워드	1	24%
5	서론	6	6%
6	본론	5	9%
7	결론	7	5%
8	참고문헌	4	11%

<표 5>에서와 같이 사용자 검색 행위를 바탕으로 한 태그의 중요 순위는 키워드, 제목, 초록, 참고문헌, 본론, 서론, 결론, 목차의 중요도를 갖는다는 것을 알 수 있다. 또한 키워드와 제목의 가중치는 거의 차이가 없는 순위를 보였다. 이와 같은 태그의 순위는 실험 대상 문서의 종류에 따라 태그가 다르기 때문에 다른 종류의 문서에서는 본 실험과 다른 결과가 나올 수 있다.

4.2.2 결정된 태그의 중요도 성능 평가

두 번째 실험은 첫 번째 실험결과로 얻어진 태그 가중치의 성능을 평가하기 위한 실험이다. 이 실험은 두 가지 측면에서 실시한다. 첫 번째는 일반적인 정보검색 평가 척도인 정확률(precision ratio)과 재현율(recall ratio)을 이용하여 평가하고, 두 번째는 문서순위결정 평가에 맞는 적합률(relevance ratio)을 이용하여 평가한다. 일반적인 정확률 공식은 (식 8)과 같고, 재현율 공식은 (식 9)와 같다[10].

$$\text{정확률} = \frac{\text{검색된 적합 문서 수}}{\text{검색된 문서 총 수}} \tag{8}$$

$$\text{재현율} = \frac{\text{검색된 적합 문서 수}}{\text{적합한 문서 총 수}} \tag{9}$$

그리고 정확률을 응용한 적합률 공식은 (식 10)과 같다[7].

$$\text{적합률} = \frac{\sum_{i=1}^{n} R_{score}}{\sum_{i=1}^{n} R_{\max}} \times 100 \tag{10}$$

단, R_{score} : 사용자가 평가한 논문의 적합성 정도로서 표현 범위는 0~3 값이고, 값에 따른 의미는 다음과 같다.

0 : 비적합, 1 : 보통
2 : 적합, 3 : 매우 적합

$R_{\max}$: 최고 적합한 정도로서 값은 3.

n : 순위가 결정된 논문의 상위 α% 내의 순위를 갖는 문서의 개수(α는 사용자가 입력하는 값)

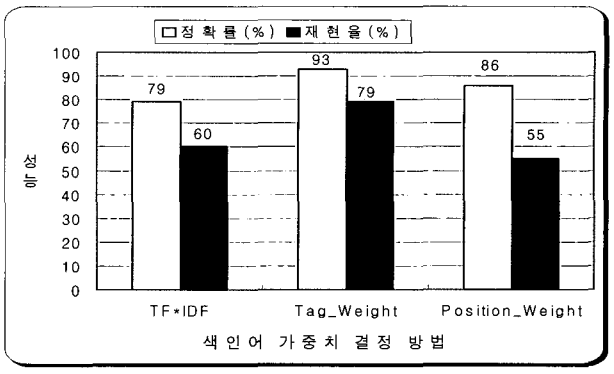
본 논문에서 제안하는 태그의 중요도가 색인어 가중치 계산에 미치는 영향과 더불어 검색성능에 미치는 영향을 알아보기 위하여 다음과 같은 방법으로 색인어 가중치를 결정한 후, 성능을 평가하고 서로 비교한다.

(1) TF*IDF : *TF*· *IDF* 방법만을 반영
(2) Tag_Weight : *TF*· *IDF* 방법+본 논문에서 제안하는 태그의 중요도 반영
(3) Position_Weight : *TF*· *IDF* 방법+문서의 단락 위치 가중치[4]를 반영

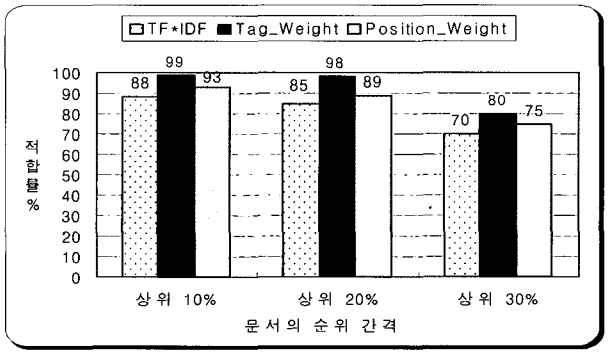
용어를 추출한 다음, 위 세 가지 방법에 의해 용어의 가중치를 계산하고, 가중치를 [0,1] 범위로 정규화한 후, 0.5 이상의 가중치를 가진 용어들을 색인어로 선정하였다. 본 논문에서는 기본적으로 *TF*· *IDF*를 이용하고 다른 방법들을 적용하고 있으므로, 색인어 선정 기준에 큰 비중을 두지 않았다. 평가자 집단은 전자정보통신 분야의 전공자 33명이 약 3개의 관심분야를 갖고 각 15회에 걸쳐서 실험을 실시하였다. 사용자가 입력한 질의와 일치하는 색인어의 가중치가 0.5이상인 문서만을 검색 결과로 했을 경우, 각 방법들을 이용한 검색결과의 평균 정확률과 평균 재현율은 (그림 2)와 같다.

태그 가중치의 성능 평가를 위한 두 번째 실험은 검색결과로 제시된 문서들을 대상으로 문서순위결정을 수행한다. 질의와 일치되는 색인어의 가중치가 0.5 이상인 문서를 검색 결과로 제시되는 실험만으로는 [0.5, 1]인 값을 갖는 색인어 가중치들의 성능을 제대로 평가할 수 없으므로 문서순위결정을 수행한다. 문서순위 결정방법은, 본 논문의 연구 범위를 벗어나며, 색인어 가중치의 성능을 평가하기 위한 실험이므로, 가장 기본적인 벡터기반 방법을 이용하였다. 질의와 일치하는 색인어의 가중치의 합을 구하여 각 검색된 문서들의 적합성 정도를 계산한다. (그림 3)은 (식 10)에서 α가 10인 경우, 20인 경우, 30인 경우 각각을 도식화한 것으로서 그래프는 관심분야별로 15회에 걸쳐 수행한 결과의 평균 적합률을 나타낸다.

"Tag_Weight"는 본 논문에서 제안하는 방법을 이용하여 색인어 가중치를 계산한 것을 뜻하고, "TF*IDF"는 *TF*· *IDF* 방법으로 색인어 가중치를 계산한 것을 뜻한다. "Position_Weight"는 [4]의 방법으로 색인어의 가중치를 계산한 것이다. "Position_Weight"의 자체 성능 평가는 모든 텍스트로부터 중요한 색인어를 추출하므로 복잡하고 시간 비용이 다소 높았다. 본 논문에서 제안하는 태그 가중치 결정 방법을 이용하면 자동색인뿐만 아니라 문서순위결정 기법의 성능 향상에 큰 도움이 될 것이다.



(그림 2) 성능 평가 - 정확률과 재현율



(그림 3) 성능 평가 - 적합률

5. 결 론

웹이 1990년대 중반부터 폭발적으로 성장해 오면서 인터넷을 이용하는 인구의 수도 급격하게 증가하게 되었으며, 웹에서 보다 효과적으로 색인을 추출하고 검색 편의성을 제공하기 위한 대안으로서 XML(eXtensible Markup Language)이 등장하였다.

XML의 태그 정보를 이용하여 검색성능을 향상시키고자, 본 논문에서는 사용자의 검색 행위를 반영하여 XML 태그 가중치를 계산하는 방법을 제안하였다. 그리고 결정된 태그의 중요도를 평가하기 위해 두 가지 실험을 실시하였다. 첫 번째 실험은 본 논문에서 제안한 방법으로 테스트 문서 집합에서 태그 가중치를 계산하여 태그의 중요 순위를 결정하는 실험이다. 태그의 중요도를 구하는 실험 결과, 태그의 중요 순위는 키워드, 제목, 초록, 참고문헌, 본론, 서론, 결론, 목차 순이었다. 또한 키워드와 제목 태그의 가중치 차이가 매우 적었다. 이 순서는 실험대상 문서의 종류에 따라 태그가 다르기 때문에 다른 부류의 문서의 경우엔 본 실험과 다른 결과가 나올 수 있다. 두 번째 실험은 태그 가중치를 반영하여 색인어 가중치를 결정한 후, 검색 성능을 평가하기 위한 실험이다. 그 결과를 살펴보면, 본 논문에서 제안하는 방법의 정확률과 재현율 그리고 적합률이 모두 좋은 성능을 나타냄을 알 수 있었다. 본 논문에서 제안하는 태그의 중요도 결정 방법을 이용하여 색인어 가중치를 결정하면 사용자

에게 보다 적합한 검색 결과를 제공할 수 있을 것이고, 문서순위결정 방법과 같이 사용되어 사용자에게 검색 편의성을 제공할 수 있을 것이다.

본 논문에서 제안하는 방법이 기술 분야 논문이 아닌 일반적인 XML 문서의 태그들에도 적용될 수 있는가에 대한 검증이 필요하다. 그리고 태그 가중치를 결정하는 과정에서 태그 가중치 벡터의 생성과 갱신 부분에 문서빈도(*DF*)와 용어빈도(*TF*)가 미치는 영향에 관하여 연구를 계속할 계획이다. 또한 질의 가중치 벡터의 가중치를 0과 1이 아닌 사용자의 선호도를 반영할 수 있는 가중치로 표현했을 경우의 효과에 대해서도 연구를 계속할 계획이다.

참 고 문 헌

[1] 고영중, 박진우, 서정연, "문장 중요도를 이용한 자동문서 범주화", 정보과학논문지 제29권 제6호, 2002.

[2] 김영란, "XML DTD의 효율적인 검색을 위한 구조 정보 및 인덱스 메카니즘", 컴퓨터정보학회 논문지 제8권 제2호, 2003.

[3] 김종영, 김철수 "가중치를 가지는 웹문서 색인기법에 관한 연구", 한국정보처리학회, 제9권, 제2호, 2002.

[4] 김홍남, 이기성, 조근식 "가중치가 부여된 규칙을 이용한 문서분류", 한국정보과학회지, 제30권, 제2-1호, pp.154~156, 2003.

[5] 박종관, 손충범, 강형일, 유재수, 이병엽, "XML 문서의 효율적인 구조 검색을 위한 색인 모델", 정보처리학회논문지D, 제8-D권 제5호, 2001.

[6] 양권묵, 박건일, 김유성, "한글 학술 논문의 일반구조를 이용한 자동 색인어 선정 시스템", 인하대학교 학위논문, 1998

[7] 우선미, 유춘식, 김용성, "용어 연관성 분석을 이용한 사용자 위주의 문서순위결정 기법", 한국정보과학회 논문지, 제28권, 제2호, pp.149-156, 2001.

[8] 유춘식, 우선미, 유철중, 이종득, 권오봉, 김용성, "자연어 처리, 통계적 기법, 적합성 검증을 이용한 자동 색인 시스템에 관한 연구", 정보처리논문지, 제5권 제6호, 1998

[9] 유춘식, "유사한 구조를 가지는 XML 문서들의 DTD 통합 알고리즘", 전북대학교 전산통계학과 박사학위논문, pp.1-108, 2005. 2

[10] 정영미, 정보검색론, 구미무역(주) 출판부, pp.1-354, 1993.

[11] 정영미, 이재윤, "지식 분류의 자동화를 위한 클러스터링 모형 연구", 정보관리학회지 제18권 제2호, 2001.

[12] 조윤기, 조정길, 이병렬, 구연설, "XML 문서에 포함된 구조 정보의 표현과 검색", 정보처리학회논문지D, 제8-D권 제4호, 2001.

[13] Anthony Hunter, "Logical Fusion rules for merging structured news reports," Data & Knowledge Engineering, Vol.42, pp.23-56, 2002.

[14] Anthony Hunter, "Merging structured text using temporal knowledge," Data & Knowledge Engineering, Vol.41, pp. 29-66, 2002.

[15] Brian Lowe, Justin Zobel, Ron Sacks-Davis "A Formal Model for Databases of Structured Text," Proceedings of the Fouth International Conference on Database Systems for Advanced Applications(Dasfaa '95), pp.449-456, 1995.

[16] Fabio Crestani, Jesus Vegas, Pablo de la Fuente, "A graphical user interface for the retrieval of hierarchically structured documents," Information rocessing and Management, Vol.40, pp.269-289, 2004.

[17] S.H.Lin, M.C.Chen, J.M.Ho, Y.M.Huang. "ACIRD : Intelligent Internet Organization and Retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol.14, No.3, May/June 2002.

[18] T. Dao, R. Sacks-Davis and J. A. Thom "An indexing scheme for structured documents and its implementation," In Proceedings of the 5th International conference on Database Systems for Advanced Applications, pp.125-134, Melbourne, Australia, April, 1997.

[19] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, and Fraçnois Yergeau, "XML 1.0(Third Edition)," W3C Recommendation, http://www.w3.org/TR/2004/REC-xml-20040204, Feb., 2004.

[20] Toung Dao "An Indexing Model for Structured Documents to Support Queries on Content, Structure and Attributes," Proceeding of ADL'98, pp.88-98, 1998.

우 선 미

e-mail : smwoo@chonbuk.ac.kr
1991년 서남대학교 전자계산학과(이학사)
1995년 전북대학교 전산통계학과(이학석사)
2001년 전북대학교 전산통계학과(이학박사)
2001년~2003년 전북대학교 중앙도서관 전산실 조교
2004년~현재 전북대학교 전북지역전자정보사업단 기금교수
관심분야 : 사용자 위주의 정보검색, 문서순위결정, 정보 필터링, XML 응용, 디지털 도서관 등

유 춘 식

e-mail : csyoo@chonbuk.ac.kr
1991년 전북대학교 전산통계학과(이학사)
1994년 전북대학교 전산통계학과(이학석사)
2005년 전북대학교 전산통계학과(이학박사)
관심분야 : XML, 스키마 통합, 유비쿼터스, 적응형 사용자 인터페이스, W3 웹 서비스 등

김 용 성

e-mail : yskim@chonbuk.ac.kr
1978년 고려대학교 수학과(이학사)
1984년 광운대학교 전산학과(이학석사)
1992년 광운대학교 전산학과(이학박사)
1985년~현재 전북대학교 전자정보공학부 교수
1996년~1998년 한국학술진흥재단 전문위원
관심분야 : XML, 사용자 중심의 정보검색, 다중 사용자 인터페이스, W3C 웹 서비스 등