DOI: 10.3745/KIPSTB.2009.16-B.1.71

XML 태그를 분류에 따른 가중치 결정

정 혜 진[†] · 김 용 성[††]

요 약

보다 효과적인 색인어 추출 및 색인어 가중치 결정을 위하여 문서의 내용뿐 아니라 구조를 이용하여 색인을 추출하는 연구가 이루어지고 있는데, 대부분의 연구들이 XML 태그의 중요도가 아닌, 문맥상의 단락에 대한 중요도를 계산하는게 일반적이다. 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 태그를 중요도에 따라 분류하고, 낮은 태그에서 추출된 용어 가중치를 계산하고, 그 가중치로 높은 가중치의 태그에서 추출된 용어의 가중치를 갱신해 가면서 최종 가중치를 계산하는 방법을 제안한다. 보다 객관적인 가중치 결정을 위하여 사용자가 중요하게 생각하는 태그를 실험해 보고 그에 따라 중요도를 분류하여 가중치 계산에 반영한다. 그리고 기존 태그 중요도 결정 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

An XML Tag Indexing Method Using on Lexical Similarity

Hye-Jin Jeong[†] · Yong-Sung Kim[††]

ABSTRACT

For more effective index extraction and index weight determination, studies of extracting indices are carried out by using document content as well as structure. However, most of studies are concentrating in calculating the importance of context rather than that of XML tag. These conventional studies determine its importance from the aspect of common sense rather than verifying that through an objective experiment. This paper, for the automatic indexing by using the tag information of XML document that has taken its place as the standard for web document management, classifies major tags of constructing a paper according to its importance and calculates the term weight extracted from the tag of low weight. By using the weight obtained, this paper proposes a method of calculating the final weight while updating the term weight extracted from the tag of high weight. In order to determine more objective weight, this paper tests the tag that user considers as important and reflects it in calculating the weight by classifying its importance according to the result. Then by comparing with the search performance while using the index weight calculated by applying a method of determining existing tag importance, it verifies effectiveness of the index weight calculated by applying the method proposed in this paper.

Keywords : XML Tag Weight, Automatic Indexing, Information Retrieval

1. 서 론

하이퍼 텍스트(Hyper Text) 개념의 등장으로 인하여 웹(World Wide Web; WWW)이 1990년대 중반부터 폭발적으로 성장해 오면서 인터넷은 빠른 속도로 전 세계 데이터 통신 구조의 표준으로 자리 잡아 가고 있다. 이에 따라 인터넷을 이용하는 인구의 수도 급격하게 증가하고 있고, 웹의 발달과 인터넷의 보편화로 인하여 자신이 원하는 정보를 얻기가 점점 어려워지고 복잡해지므로 웹에서 보다 효과적으로 색인을 추출하고 검색 편의성을 제공하는 연구가 필요하다. 이러한 문제점을 해결하기 위한 대안 중의 하나가 정보를 XML(eXtended 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 인식되고 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 검색할 수 있는 기능도 필요하다[1]. 또한 문서의 내용 정보뿐만 아니라 문서의 구조 정보를 이용하여 색인을 추출하고 색인어 가중치를 계산할

† 준 회 원 : 전북대학교 컴퓨터정보학과 박사수료
†† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2007년 12월 12일
수 정 일 : 1차 2008년 11월 10일
심사완료 : 2008년 11월 30일

수 있다면 검색 효율성을 높일 수 있을 것이다. 태그의 중요도에 관한 연구가 일부에서 이루어지고 있다[2,3]. 하지만, 객관적인 실험을 통해서 중요도를 입증하기보다는 상식적인 관점이나 전문가의 휴리스틱(heuristic)에 의하여 단순한 수치로 중요도를 결정하고 있다. 따라서 본 논문에서는 XML 태그를 이용해 추출된 색인의 위치정보를 이용해서 위치별 색인어의 가중치를 계산하는 기법을 제안하고, 본 논문에서 제안한 방법을 적용하여 색인어 가중치를 이용한 검색성능과 비교함으로써 계산된 색인어 가중치의 효과를 검증한다.

실험 평가는 웹에서 검색한 컴퓨터과학 및 정보통신 분야의 논문과 기사들을 이용한다. XML로 표기되지 않는 것들은 XML로 재구성하여 사용한다.

먼저 2장에서 XML 문서의 구조 및 태그의 종류를 간단히 살펴보고 문서 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 위해 실험하고, 마지막으로 5장에서 결론과 함께 향후 연구 방향에 관해 기술한다.

2. 관련 연구

단어에 가중치를 부여하는 목적은 한 문서가 취급하고 있는 개념들의 주제적 요소로서의 중요도에 따라 색인어로서 상대적 가치를 표현하기 위함이다. 자동색인 기법에서 색인어 가중치 결정은 주로 통계적 기법을 이용하는데, 통계적 기법의 통계적 기준은 모두 단어의 출현빈도에 근거하고 있다. 단어빈도를 문헌빈도로 나누어주는 역문헌빈도(tf*idf)에 의한 색인어 후보의 가중치 기법과 표제 색인어 후보의 가중치 기법 등이 많이 사용되고 있다. 이러한 방법과 XML 태그의 중요도를 이용한 방법을 다음 소절에서 설명한다.

2.1 문서의 부분 중요도를 이용한 색인어 결정 방법

색인어의 가중치를 계산할 때 색인어가 가중치가 부여된 문서의 위치 즉, 제목, 초록, 키워드, 서론, 관련연구, 내용, 실험, 결론, 감사의 글, 참고문헌에서 키워드 빈도수를 이용한 가중치 계산[4]이나 중요도가 높은 태그 위치에 따라 가중치를 계산[5]하는 방법으로는 최종 색인어를 결정할 수 있다.

2.1.1 가중치가 부여된 단어

일반적으로 문서의 “제목”이나 “키워드” 위치에서 추출된 키워드는 “서론”이나 “관련연구”의 위치에서 추출된 키워드보다 더 중요도를 가지고 있다는 가설하에 키워드의 빈도수(Term Frequency)와 <표 1>과 같이 문서의 위치에 부여된 가중치(Weight)로 문서에 대한 단어의 지지도(Term Support)를 측정하여 문서에 대한 키워드의 중요도가 높은 키워드들로 연관 규칙을 적용하고 있다.

문서의 각 단락에서 추출된 키워드 t_i는 (식 2-1)과 같이 키워드의 빈도수와 문서 위치에 부여된 가중치를 이용하여 계산한다.

〈표 1〉 문서의 위치에 따른 가중치

	가중치	비 고
제목	1.9	
초록	1.8	
키워드	2	
서론	1.3	
관련연구	1.6	
내용	1.5	
실험	1.2	
결론	1.2	
감사의 글	1	
참고문헌	1.4	

$$Sup_{ti} = \frac{sup'_{ti}}{MAX\{sup'_{ti}\}} \cdots \text{(식 2-1)}$$

단 $Sup'_{ti} = \sum_{Sj} tf_{ij} \cdot W_{sj}$ 에 의하여 계산한다.

tf_{ij}는 문서의 위치 S_j에 있는 t_i의 빈도수를 나타내며 W_{sj}는 문서의 위치 가중치를 나타낸다.

2.1.2 태그 정보를 이용한 가중치 부여하는 방법

웹문서는 태그(tag)로 이루어진다는 특징에 착안[4-6]하여 색인어로서 중요한 내용을 담고 있는 태그와 중요하지 않는 태그를 추출하여 <표 2>와 같이 1~9의 범위를 갖는 가중치

〈표 2〉 태그 가중치 순위

순	태그	순	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 요
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

값을 font size가 1~3과 <h5>, <h6>은 기준 가중치 1로 하고 링크가 된 곳을 최고의 가중치 9점을 주었다. 다음으로 제목, 주제, 효과, 강조로 나누어 2~8까지의 가중치를 주었다.

태그가 2개 이상이 중복될 경우에는 <표 3>의 가중치 테이블을 기준으로 하여 상위 가중치를 가진 태그에만 가중치를 계산하였다.

2.1.3 태그별 색인어 가중치 결정 기법

XML 문서의 자동색인 및 색인어 가중치 결정을 위한 태그의 가중치를 계산하기 위하여, 일정한 XML 테스트 문서 집단(논문이나 연구 보고서)을 만들어 사용자의 검색 행위를 알아보고, XML 문서에서 색인어를 추출하여 주요 태그마다 태그 용어 벡터와 태그 가중치 벡터를 생성한다[2,3].

[3]은 태그가중치의 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 $T_i = (t_{i1}, t_{i2}, ..., t_{in})$과 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 초기 값은 $TF \cdot IDF$ 방법으로 구하는데, 해당 문서에서 출현하지 않은 용어에 대응되는 가중치는 0이다. 그런 후 문서마다 출현하는 용어를 8개의 태그 용어 벡터와 비교하여 일치하는 용어가 발생하면, 태그 가중치인 V_tag_j를 그 용어의 가중치에 반영해 준다. 태그가중치를 반영하여 최종 용어의 가중치를 계산하는 공식은 (식 2-2)와 같다.

$$W_{ik} = \sum_{j=1}^{j=8} |W_{ik}(1 + V\_tag_j)| \cdots\cdots\cdots\cdots \text{ (식 2-2)}$$

단, W_{ik} : i번째 문서의 k번째의 용어의 가중치

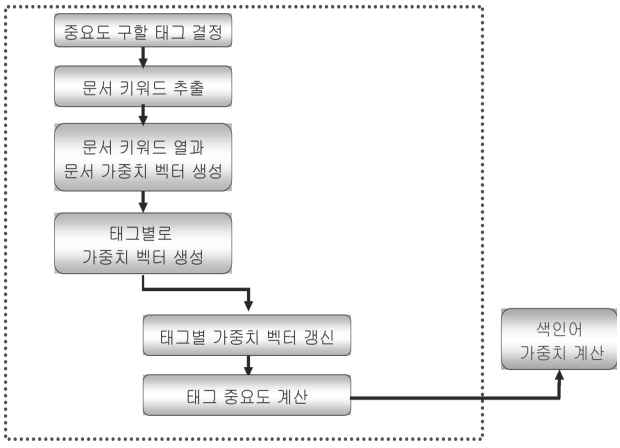
V_tag_j : W_{ik}이 용어가 포함되어 있는 j번째 태그가중치

3. XML 태그 가중치를 이용한 색인어 가중치 계산

본 장에서는 학위 논문을 대상으로 XML 문서 태그의 가중치를 계산하고, 이 태그의 중요도를 이용하여 색인어 가중치를 계산하는 방법을 기술한다. XML 태그의 가중치를 결정하는 과정은 (그림 1)과 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>과 같다.

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. 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.



(그림 1) 태그별 색인어 가중치 결정 흐름도

〈표 4〉 설문조사로 선정한 XML(논문) 태그

비주제 색인		주제 색인	
No.	tag	No.	tag
1	저자	1	제목
2	출판년도	2	목차 (그림, 표, 수식 포함)
3	출처	3	초록
		4	키워드
		5	서론
		6	본론 (관련연구, 연구내용 포함)
		7	결론
		8	참고문헌

3.1 태그별 색인어 가중치 결정 기법

일반적으로 논문을 검색할 때, 제목, 키워드, 초록을 가장 먼저 검색하게 된다. 그것은 논문의 제목이나 키워드, 초록에 그 논문을 대표할 수 있는 용어가 크게 비중을 두고 있기 때문이다. 따라서 본 논문에서는 주제 색인을 제목, 키워드, 초록과 같이 사용자 검색을 우선으로 하는 태그와, 논문을 대표할 수 없는 용어가 빈번히 발생할 수 있는 확률이 많아 중요도가 낮은 태그, 그외 태그를 <표 5>와 같이 분류하여 태그 가중치를 결정하였다. 그리고 그 분류에 따라 중요도가 낮은 태그부터 가중치를 계산한 후 중간태그에 반영하여 가중치를 갱신하였다. 그 후 중요도가 높은 태그에 반영하여 최종 가중치를 갱신하도록 하였다. 자세한 방법은 다음 소절에서 설명한다.

3.1.1 문서 용어열과 문서 가중치 벡터 생성

태그의 중요도 계산에 사용하고 색인어 선정에 사용하기 위하여 문서집합에서 추출한 용어로 문서 용어열 $T_doc = (dt_1, dt_2, ..., dt_n)$과 문서 가중치 벡터 $W_doc = (dw_1, dw_2, ..., dw_n)$을 생성한다. 이때 n은 XML 문서 집합을 구성하는 문서에서 추출한 키워드의 수이다. 문서 가중치 벡터는 문서 용어열과 쌍을 이루는 벡터로서 각 용어에 대한 가중치를 나타낸다. 문서 가중치 벡터의 값은 문서의 용어 각각에 대한 가중치로서 역문헌빈도($TF \cdot IDF$) 방법을 이용한다.

이때 얻어진 가중치는 중요도가 낮은 태그의 가중치로 한다.

가중치 벡터는 태그별 용어열 $Ti_tag = (tag_i_t_1, tag_i_t_2, ..., tag_i_t_n)$과 가중치 벡터 $Wi_tag = (i_tw_1, i_tw_2, ..., i_tw_n)$을 생성한다. 이때 i는 중요도에 따라 중요도가 낮은 태그인경우 1, 중간인 경우는 2, 높은 경우는 3의 값을 갖는다.

3.1.2 중요도가 낮은 태그 용어열과 가중치 벡터 생성

중요도가 낮은 태그 용어열 $T1_tag = (tag1_t_{j1}, tag1_t_{j2}, ..., tag1_t_{jn})$과 가중치 벡터 $W1_tag = (1_tw_{j1}, 1_tw_{j2}, ..., 1_tw_{jn})$을 생성한다.

중요도가 낮은 태그 용어열은 문서 집합의 모든 문서의 해당 태그에 속한 용어로 구성한다. 예를 들어 본문 태그 용어열 $T1_tag$는 모든 문서들의 전체에 포함된 용어들로 구성된다. 즉, $T1_tag = (1_tag_t_{11}, 1_tag_t_{12}, ..., 1_tag_t_{1n})$이고, n은 모든 문서들의 제목에 포함된 용어의 수이다. 이때 벡터의 크기는 문서 용어열 T_doc와 같다.

각 태그별 용어열에 대응되는 가중치를 나타내는 태그별 가중치 벡터 $W1_tag$는 $T1_tag$에 대응되는 가중치 벡터이다. 예를 들어 본문 가중치 벡터 $W1_tag$는 $W1_tag$과 쌍을 이루는 가중치를 표현한다.

중요도가 낮은 태그의 가중치 벡터는 문서 용어열의 용어를 포함하지 않는 부분은 0값을 가지고, 문서 용어열의 용어를 포함하는 부분은 (식 3-1)로 계산하여 태그별 용어 가중치 벡터 $W1_tag_k$를 생성한다. 이때 k는 중요도가 낮은 태그(tag)의 k번째 용어 가중치를 의미하는 것으로써 문헌빈도를 반영하기 위해 $W1_tag_k$의 값은 W_doc의 dw_k에 $T1_tag_k$가 출현하는 문서의 개수(문헌빈도)를 곱해준 값이다. 이때 $T1_tag$는 중요도가 낮은 태그에 포함된 용어열이다.

$$W1\_tag_k = \mid W\_doc_k * DF\_T1\_tag_k \mid \quad \cdots\cdots\cdots\cdots \text{(식 3-1)}$$

단, W_doc_k : W_doc의 k번째 용어 가중치

$DF\_T1\_tag_k$: 낮은 중요도로 분류된 태그에서 k번째 용어가 출현한 문헌빈도

3.1.3 중요도가 중간인 태그의 가중치를 반영한 가중치 벡터의 갱신

중요도가 낮은 태그의 가중치 벡터가 생성되면, 태그별 가중치 벡터와 중요도 태그에 따른 태그 벡터를 비교하고 (식 3-2)에 의해 태그별 가중치 벡터의 값을 갱신한다.

중요도가 낮은 태그 용어열의 용어를 포함하지 않는 부분은 1값을 가지고, 포함하는 경우에는(식 3-2)으로 계산하여 중요도가 중간인 태그별 용어 가중치 벡터 $W2_tag_k$를 생성한다. 이때 k는 중요도가 중간인 태그로 분류된 태그(tag)의 k번째 용어 가중치를 의미하는 것으로써 문헌빈도를 반영하

〈표 5〉 중요색인 분류 태그

주제 색인					
No.	중요도가 낮은 태그(1)	No.	중요도가 중간인 태그(2)	No.	중요도가 높은 태그(3)
1	본론 (관련연구, 연구내용 포함)	1	목차 (그림, 표, 수식 포함)	1	제목
		2	서론	2	초록
		3	결론	3	키워드
		4	참고문헌		

기 위해 $W2_tag_k$의 값은 $W1_tag_k$의 dw_k에 $T2_tag_k$가 출현하는 문서의 개수(문헌빈도)를 더해준 값이다. 이때 $T2\_tag$ 는 중요도가 중간이 태그에 포함된 용어열이다.

$$W2\_tag_k = | W1\_tag_k + DF\_T2\_tag_k | \cdots\cdots\cdots \text{(식 3-2)}$$

단, $W1\_tag_k$: 중요도가 낮은 태그중간인 태그의 k번째 용어 가중치

$DF\_T2\_tag_k$: 중간 중요도로 분류된 태그에서 k번째 용어가 출현한 문헌빈도

3.1.4. 중요도가 높은 태그 가중치를 반영한 가중치 벡터의 갱신

3.1.3에 의해 갱신된 가중치는 다시 가중치가 높은 태그별 가중치 벡터와 중요도 태그에 따른 태그 벡터를 비교하고 (식 3-3)에 의해 태그별 가중치 벡터의 값을 갱신한다.

중요도가 높은 태그 용어열의 용어를 포함하지 않는 부분은 1값을 가지고, 포함하는 경우에는(식 3-3)으로 계산하여 중요도가 중간인 태그별 용어 가중치 벡터 W_tag_k를 생성한다. 이때 k는 중요도가 높은 태그로 분류된 태그(tag)의 k번째 용어 가중치를 의미하는 것으로써 문헌빈도를 반영하기 위해 W_tag_k의 값은 $W2_tag_k$의 dw_k에 $T3_tag_k$가 출현하는 문서의 개수(문헌빈도)를 더해준 값이다. 이때 $T3\_tag$ 는 중요도가 중간이 태그에 포함된 용어열이다.

$$W\_tag_k = | W2\_tag_k + DF\_T3\_tag_k | \cdots\cdots\cdots \text{(식 3-3)}$$

단, $W2\_tag_k$: 중요도가 낮은 태그중간인 태그의 k번째 용어 가중치

$DF\_T3\_tag_k$: 높은 중요도로 분류된 태그에서 k번째 용어가 출현한 문헌빈도

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

$$N\_W\_tag_k = \frac{W\_tag_k - W_{\min}}{W_{\max} - W_{\min}} \cdots\cdots\cdots\cdots \text{(식 7)}$$

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

$TF \cdot IDF$ 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 태그가중치를 반영하여 계산하면 최종 용어의 가중치는 [0, 1] 범위를 넘을 수도 있다. (식 7)에 의해 용어의 가중치를 정규화 시킴으로써 [0, 1] 범위의 값을 갖는 각 문서에 대한 최종 색인어 가중치 $N_W_tag_k$를 계산한다.

4. 실험 및 평가

본 장에서는 본 논문에서 제안하는 방법에 의하여 태그의 가중치를 결정하기 위한 실험과, 결정된 태그 가중치를 이용한 검색 성능을 평가한다. 검색 성능은 결정된 태그 가중치를 반영하여 색인어 가중치를 결정한 후, 일반적인 검색 성능 평가 척도를 이용하여 성능을 평가한다. 또한 본 논문에서 제안하는 태그의 가중치를 반영하여 계산된 색인어 가중치의 성능을 확신하기 위하여 문서 순위 결정을 수행한 후, 상위 문서의 적합성 정도를 평가한다. 본 논문에서는 실험 평가를 위하여 정확률(precision ratio)과 재현율(recall ratio), 그리고 적합률(relevance ratio)[7]을 이용하였다.

4.1 실험 환경

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

실험 데이터
- 웹에서 검색한 컴퓨터과학 및 정보통신 분야의 논문과 기사를 XML로 재구성한 데이터 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

크게 두 가지 측면에서 실험평가를 실시한다. 첫 번째 실험 평가는 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 KT-Set의 문서집단에서 5개의 관심분야별로 각각 20명씩 전공자로 하여금 평균 15회 이상 질의를 입력하여 검색을 실시하게 하였다. 이 때, 한사람이 여러 관심분야를 가질 수 있도록 하였다. 총 참여 인원은 33명이고, 33명이 평균 3개의 관심분야를 테스트하게 하여 질의 총 횟수는 약 900이다. 그리고 본 논문에서 제안하는 과정에 의해 계산하여 각 분야별로 평균값을 구하면 <표 6>과 같다.

<표 6>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 키

〈표 6〉 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%

워드와 제목의 가중치는 거의 차이가 없는 순위를 보였다. 이 순서는 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 실험 평가는 첫 번째 실험결과로 얻어진 태그 가중치의 성능을 평가를 위한 실험이다. 이 실험은 두 가지 측면에서 실시한다. 첫 번째는 일반 정보검색 평가 척도인 정확률(precision ratio)과 재현율(recall ratio)을 이용하여 평가하고, 두 번째는 문서순위결정 평가에 맞는 적합률(relevance ratio)을 이용하여 평가한다.

$$\text{정확률} = \frac{\text{검색된 적합 문서 수}}{\text{검색된 문서 총 수}} \quad \cdots\cdots\cdots\cdots \text{(식 8)}$$

$$\text{재현률} = \frac{\text{검색된 적합 문서 수}}{\text{적합한 문서 총 수}} \quad \cdots\cdots\cdots\cdots \text{(식 9)}$$

그리고 정확률을 응용한 적합률 공식은 (식 10)과 같다[9].

본 논문에서 제안하는 태그의 중요도가 색인어 가중치를 계산에 미치는 영향과 더불어 검색성능에 미치는 영향을 알아보기 위하여 다음과 같은 방법으로 색인어 가중치를 결정할 수 검색 성능을 평가하고 서로 비교한다.

$$\text{적합률} = \frac{\sum_{i=1}^{n} R_{score}}{\sum_{i=1}^{n} R_{\max}} \times 100 \quad \cdots\cdots\cdots\cdots \text{(식 10)}$$

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)

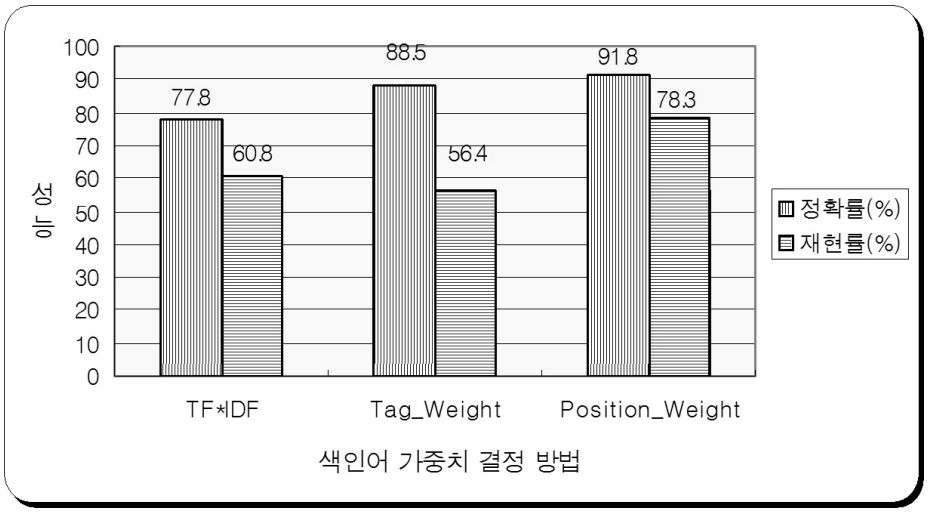
(1) TF*IDF : $TF \cdot IDF$ 방법만을 반영

(2) Tag_Weight : $TF \cdot IDF$ 방법 + 본 논문에서 제안하는 태그의 중요도 반영

(3) Position_Weight : $TF \cdot IDF$ 방법 + 문서의 단락 위치 가중치[4]를 반영

(4) Q_Tag Weight : [3]이 제한한 사용자 질의를 반영한 태그별 가중치를 반영

용어 추출 후, 위 세 가지 방법에 의해 용어의 가중치를 계산한 후, 가중치를 [0, 1] 범위로 정규화한 후, 0.5 이상의 가중치를 가진 용어들을 색인어로 선정하였다. 본 논문에서는 기본적으로 $TF \cdot IDF$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



(그림 2) 성능 평가 - 정확률과 재현률

검색결과 평균 정확률과 평균 재현률은 (그림 2)와 같다.

"Tag_Weight"는 본 논문에서 제안하는 방법으로 색인어 가중치를 계산한 것을 뜻하고, "TF*IDF"는 $TF \cdot IDF$ 방법으로 색인어 가중치를 계산한 것을 뜻한다. "Position_Weight"는 [5]의 방법으로 색인어의 가중치를 계산한 것이다. "Position_Weight"의 자체 성능 평가는 모든 텍스트로부터 중요한 색인어를 추출하는데 복잡하고 시간 비용이 다소 높았다. 본 논문에서 제안하는 태그가중치 결정 방법을 이용하면 자동색인 뿐만 아니라 검색·문서 순위 결정 기법의 성능 향상에 큰 도움이 될 것이다.

5. 결 론

폭발적으로 성장해 오면서 인터넷을 이용하는 인구의 수도 급격하게 증가하게 되었다. 사용자가 원하는 정보의 양 또한 기하급수적으로 증가하여 적합한 정보를 얻기가 점점 어려워지고 있다. 따라서 웹상에서 보다 효과적으로 색인을 추출하고 검색 편의성을 제공하기 위한 대안으로서 XML (Extended Markup Language)이 등장하였다.

현재 XML 문서의 구조적 정보를 이용하여 검색 효율을 높이기 위한 연구가 활발히 진행되고 있다.

본 논문에서는 XML을 구성하고 있는 태그를 분석하고 태그별로 중요도를 달리하여 분류하였다.

중요도가 가장 낮은 태그의 용어 가중치를 먼저 계산 후 중요가 중간인 태그에 포함된 용어 가중치와 더해줌으로써 가중치를 갱신하였다. 갱신된 가중치는 다시 한번 가장 중요한 태그에 포함된 영어 가중치와 더해줌으로써 최종 가중치를 갱신하였다.

이 실험을 위해 두 가지 실험을 실시하였다.

첫 번째 실험은 사용자에게 설문을 통하여 논문을 구성하는 태그들 중에서 중요도를 구할 태그를『제목』,『목차』,『초록』,『키워드』,『서론』,『본론』,『결론』,『참고문헌』으로 선정하였다. 두번째 실험은 문서 전체에서 추출한 용어로 문서 용어열 T_doc를 구성하고, $TF \cdot IDF$ 방법으로 T_doc에 대응되는 문서 가중치 벡터 W_doc를 구성하였다. 각 태그마다 태그 용어열 Ti_tag_k와 이에 대응되는 태그 가중치 벡터 Wi_tag_k를 구성하여 W_tag_k를 갱신하고 최종 태그의 가중치를 계산하였다.

태그의 중요도를 구하는 실험 결과, 태그의 중요 순위는『키워드』,『제목』,『초록』,『참고문헌』,『본론』,『서론』,『결론』,『목차』 순이었다. 또한『키워드』와『제목』 태그의 가중치 차이가 매우 적었다.

태그의 중요도를 반영하여 색인어 가중치를 결정한 후 검색 성능을 평가해 본 결과, 중요도에 따른 태그를 분류하여 색인어 가중치를 결정하면, 사용에게 보다 적합한 검색 결과를 제공할 수 있을 것이고, 문서순위결정 방법과 같이 사용되어 사용자에게 검색 편의성을 제공할 수 있을 것이다.

앞으로 복잡한 가중치 계산으로 인한 연산 시간에 미치는 영향과 사용자의 선호도를반영할 수 있는 가중치로 표현하기 위한 방법에 관하여도 연구를 계속할 것이다.

참 고 문 헌

[1] Brian Lowe, Justin Zobel and 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.

[2] 우선미, "사용자 질의를 이용한 XML 태그의 가중치 결정", 정보처리논문지 D(정보처리 응용), 2005.

[3] 정혜진, "사용자 질의를 이용한 XML 태그이 중요도 결정 기법", 전북대학교석사학위논문, 2004.

[4] 김홍남, 이기성, 조근식 "가중치가 부여된 규칙을 이용한 문서 분류", 한국정보과학회지, 제30권, 제2-1호, pp.0154-0156, 2003.

[5] 김종영, 김철수 "가중치를 가지는 웹문서 색인기법에 관한 연구", 한국정보처리학회, 제09권, 제02호, pp.0000-0000, 2002.

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

[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", 한국정보처리학회 논문지, 제5권 제6호, 1998.

정 혜 진

e-mail : hi-jin@hanmail.net

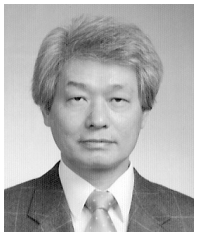
1997년 익산대학교 컴퓨터과학(전문공학사)

1999년 한국방송통신대학교 컴퓨터과학과(이학사)

2004년 전북대학교 대학원 컴퓨터정보학과(공학석사)

2007년 전북대학교 대학원 컴퓨터정보학과(공학박사수료)

관심분야 : 인공지능, 정보검색, 데이터마이닝, 멀티미디어 등



김 용 성

e-mail : yskim@chonbuk.ac.kr

1978년 고려대학교 수학과(이학사)

1984년 광운대학교 대학원 전산학과(이학석사)

1992년 광운대학교 대학원 전산학과(공학박사)

1985년~현 재 전북대학교 전자정보공학부 교수

관심분야 : XML, 정보검색, 웹서비스 등