XML 태그 분류에 따른 가중치 결정

정혜진*
전북대학교 공과대학 전자정보공학부
e-mail : arayesO@chonbuk.ac.kr

The eight decision which it follows in XML tag classification

Hye-Jin Jeong*
*Division of electronics and information engineering, Chonbuk National University

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.

1. 서 론

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. XML은 논리적 구조를 나타내는 여러 DTD를 사용할 수 있다. XML 문서는 하나의 문서에 내용정보와 구조정보를 가지고 있기 때문에 기존의 내용정보에 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 수 있다면 검색 효율성을 높일 수 있을 것이다. 태그의 중요도에 관한 연구가 일부에서 이루어지고 있다[3][4]. 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 제안한다.

2. 관련연구

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 문

헌빈도로 나누어주는 역문헌빈도(tf*idf)에 의한 색인어 후보의 가중치 기법과 표제 색인어 후보의 가중치 기법 등이 많이 사용되고 있다.

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.

(1) 가중치가 부여된 단어

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.

<표 1> 문서의 위치에 따른 가중치

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

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

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

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) 태그 정보를 이용한 가중치 부여하는 방법

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

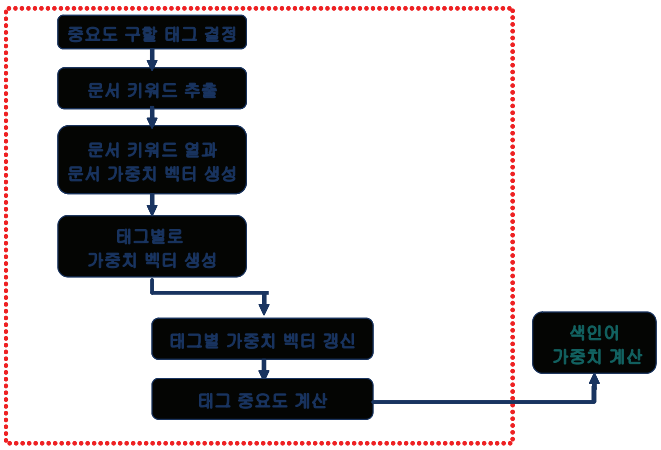
태그가 2개 이상이 중복될 경우에는 <표 2>의 가중치 테이블을 기준으로 하여 상위 가중치를 가진 태그에만 가중치를 계산하였다.

<표 2> 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

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)의 박스 안과 같다.



(그림 1) 태그별 색인어 가중치 결정 흐름도

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 있었다.

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, 그 외 태그를 <표 3>과 같이 분류하여 태그 가중치를 결정하였다. 그리고 그 분류에 따라 중요도가 낮은 태그부터 가중치를 계산한 후 중간태그에 반영하여 가중치를 갱신하였다.

<표 3> 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		

(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.

(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.

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 부분을 계산하여 태그별 용어 가중치 벡터 $W1_tag_k$를 생성한다.

(3) 중요도가 중간인 태그의 가중치를 반영한 가중치 벡터의 갱신

중요도가 낮은 태그의 가중치 벡터가 생성되면, 태그별 가중치 벡터와 중요도 태그에 따른 태그 벡터를 비교하고 태그별 가중치 벡터의 값을 갱신한다. 중요도가 낮은 태그 용어열의 용어를 포함하지 않는 부분은 1값을 가지고, 포함하는 경우에는 중요도가 중간인 태그별 용어 가중치 벡터 $W2_tag_k$를 생성한다.

(4) 중요도가 높은 태그 가중치를 반영한 가중치 벡터의 갱신

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

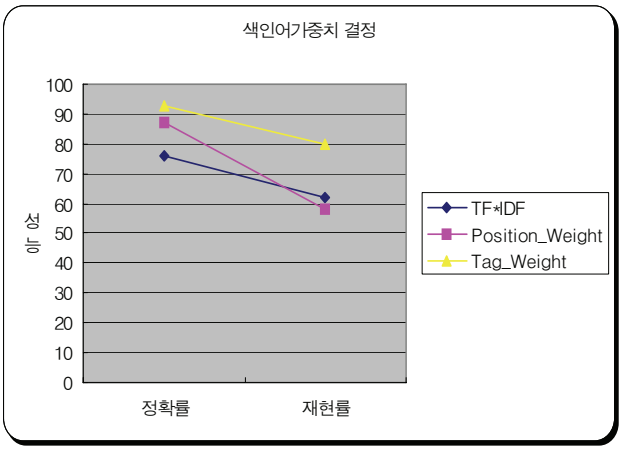
이때, 용어가 여러 태그에 중복 위치한 경우 각 태그의 중요도 값인 V_tag_j는 모두 더해지므로 색인어 가중치가 상대적으로 높아진다.

$TF \cdot IDF$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 태그가중치를 반영하여 계산하면 최종 용어의 가중치는 [0, 1] 범위를 넘을 수도 있다. 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

실험 집단은 KT-Set 문서을 XML로 변형한 논문들이고, KT-Set의 테스트 질의를 이용하여 검색 성능을 평가한다. KT-Set의 테스트 질의는 YCPARK에 의해 1995년에 만들어졌고, 총 50개의 테스트 질의들 중에서 33번 테스트 질의를 이용하여 실험 평가한다. 33번 테스트 질의는 “초고속& 정보& 통신망”이고, 결과 문서의 개수는 77개이다. 평가자 집단은 전자정보통신 분야의 전공자 30명이 각 10회에 걸쳐서 실험을 실시하였다.

사용자가 입력한 질의와 매칭되는 색인어의 가중치가 0.5이상인 문서만을 검색 결과로 했을 경우, 각 연구들의 이용한 정확률과 재현률은 (그림 2)와 같다.



(그림 2) 성능 평가 - 정확률과 재현률

태그 가중치의 성능 평가를 위한 두 번째 실험으로서 검색결과로 제시된 문서들을 대상으로 순위결정을 수행한다. 질의와 매칭되는 색인어의 가중치가 0.5이상인 문서를 검색 결과로 제시되는 실험만으로는 0.5 이상인 색인어 가중치의 성능을 제대로 평가할 수 없으므로 문서순위결정을 수행한다. 질의와 매칭되는 색인어의 가중치의 합을 구하여 각 검색된 문서들의 적합성 정도를 계산한다. 문서순위 결정 방법은 본 논문의 연구 범위가 아니고, 색인어 가중치의 성능을 평가하기 위한 실험이므로 기존에 연구된 기본적인 벡터 기반 방법을 이용하였다.

"Tag_Weight"는 본 논문에서 제안하는 방법으로 색인어 가중치를 계산한 것을 뜻하고[5], "TF*IDF"는 $TF \cdot IDF$ 방법으로 색인어 가중치를 계산한 것을 뜻한다[6]. "Position_Weight"는 [4]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. 결론 및 향후 연구 과제

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, 문서 순위결정 방법과 같이 사용되어 사용자에게 검색 편의성을 제공할 수 있을 것이다.

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

참고문헌

[1]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.

[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}, 한국 정보 과학회지, 제 30권, 제 2-1호 pp. 0154~ 0156 2003

[5] 김종영, 김철수 "가중치를 가지는 웹문서 색인기법에 관한 연구", 한국정보처리학회, 제 09권, 제 02호 pp. 0000~0000, 2002