제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

가중치를 가지는 웹문서 색인기법에 관한 연구

김종영°, 김철수
운천초등학교, 서남대학교
e-mail:namkim11@hanmail.net, chskim@tiger.seonam.ac.kr

A Study of Indexing Methods with weight-value of Web document

Jong-Young Kim, Cheol-su Kim
Unchon elementary school, Seonam University

요약

검색된 문헌들에 대한 항해 시간을 줄이기 위해서 검색된 문헌들의 문헌 순위화가 필수적이다. 문헌 순위화를 위해서는 문헌 순위화를 위한 순위화 정보가 필요하다. 본 논문에서는 검색된 문헌들에 대한 순위화를 보다 효율적으로 수행하기 위한 정보를 제공하기 위하여 HTML 문서에 대한 색인 과정에서 다양한 가중치를 가지는 색이어 추출 방법에 관하여 연구하였다.

웹문서들은 태그로 이루어지며 중요한 색인어들은 특정 태그 속에 포함되어 있다는 것에 착안하여 색인어의 중요도에 영향을 줄 수 있는 태그를 선별하고, 선별된 태그들에 대해 휴리스틱 정보를 이용하여 중요도를 부여한 후 선별된 태그에 영향을 받는 문장들에서 추출된 색인어에 대하여 가중치를 부여하는 방법을 이용하였다. 색인어 추출을 형태소 분석기를 이용하였다. 색인어들이 다양한 가중치를 가지므로 검색 과정에서 검색된 문헌들에 대하여 효율적인 순위화가 가능하여 관련 문헌을 보다 빠르게 찾을 수 있는 환경을 제공할 수 있다.

1. 서론

검색시스템의 주요 성능 평가는 일반적으로 재현율과 정확율에 평가된다[4]. 그러나 웹 문서의 급증으로 검색된 문헌 수가 지나치게 많은 경우가 대부분이다. 이 때 재현율 못지 않게 관심이 많은 문헌들을 관심이 적은 문헌들보다 먼저 볼 수 있도록 하는 문헌순위화가 검색 시간 관점에서 매우 중요하다.

문헌 순위화를 위해서는 검색된 문헌들에 대하여 중요도 차이를 비교할 수 있는 비교 대상이 필요하다. 이를 위해서 색인어들에 대한 중요도 정도를 나타내는 가중치 정보 즉, 가중치를 가지는 색인 시스템이 필요하다.

본 논문에서는 정보 검색 시스템에서 문헌 순위화를 보다 효율적으로 수행하기 위하여 색인어들이 다양한 가중치를 가지는 색인 기법에 대하여 연구하였다. 본 논문에서 HTML 문서를 대상으로 하는 색인 시스템에서 웹문서들은 태그로 이루어지며 중요한 색인어들은 특정 태그 속에 포함되어 있다는 것에 착안하여 색인어의 중요도에 영향을 줄 수 있는 태그를 선별하고, 선별된 태그들에 대해 휴리스틱 정보를 이용하여 중요도를 부여한 후 선별된 태그에 영향을 받는 문장들에서 추출된 색인어에 대하여 가중치를 부여하는 방법을 이용하였다. 색인어 추출을 형태소 분석기를 이용하였다.

2. 관련 연구

색인어의 가중치를 부여하기 위하여 통계적 방법과 자연어 처리 기법을 이용한 방법들이 제안되었다. 통계적인 방법을 이용하는 경우 대부분 색인어의 빈도수나 역문헌 빈도 정보를 이용하여 색인어 가중치를 부여할 수 있으나 문헌 순위화에는 큰 영향을 주지 않은 것으로 나타났다[1]. 보다 효율적인 색인어 가중치를 부여하기 위하여 자연어처리 기법을 이용하거나 문헌 구조를 이용하는 방법을 이용하는 방법들이 제안되었다[2].

문헌 구조를 이용한 방법[2]은 전방내용(표제, 저자, 요약), 본문(서론, 결론, 결론) 부속물(참고문헌,

부록, 저자실명)의 정보와 색인어 빈도 정보를 이용하여 색인어의 가중치를 부여한 방법으로 성능을 향상시켰다.

문헌 순위화에 관한 또다른 연구로 자연어 처리 기법과 통계적 방법을 다양하게 결합하여 문헌 순위화 기법에 대한 연구[3]에서는 보다 향상된 문헌 순위화를 수행하기 위하여 자연어 처리 기법과 문헌의 구조 정보를 이용하여 색인어들이 나타난 위치에 따라 색인어의 가중치를 부여(표제어의 색인어: 0.8, 요약의 색인어: 0.5, 표제어와 요약에 동시 출현한 색인어: 1.0)하여 문헌 순위화를 시도하였다. 이 연구에서는 기존의 통계적 방법과 색이어의 가중치 정보를 이용한 문헌 순위화의 다양한 실험을 통하여 통계적 방법만을 다양하게 결합한 기존의 방법들보다 색인어들이 출현한 위치 정보를 이용하여 색인어 가중치를 부여한 방법이 문헌 순위화에 크게 영향을 주었다.

즉, 통계적 방법보다는 자연어 처리 방법을 이용한 용어 가중치를 가지는 색인어가 문헌 순위화에 도움을 주었다.

3. 가중치를 가진 색인어 추출 방법

본 연구에서는 웹상에서 HTML 문서는 태그(tag)로 이루어진다는 특징에 착안하여 HTML 문서상에서 색인어로서 중요한 내용들을 담고 있는 태그들을 추출 분석하여 그 태그들에게 중요도 가중치를 부여하였다. 그리고 색인작업과정에 그 태그들에게 부여한 가중치를 저장하여 가중치 정보를 가진 색인어를 저장하여 두고 검색결과에 대한 문헌순위화 과정에서 색인어의 가중치 정보를 이용하여 순위화를 수행함으로써 보다 효율적인 순위화를 수행할 수 있도록 하기위해 가중치를 가지는 색인어를 추출하는 과정을 거친다.

3.1 가중치 관련 태그 추출

웹브라우저 상에 나타나는 문자에 대한 태그를 분석하여 그에 따르는 태그에 가중치를 주기 위하여 문자의 색상을 제외한 굵기, 크기, 표제어에 대한 태그를 추출하고 다시 색상을 표현하는 태그를 제외시키고 직접적으로 중복되지 않는 태그만을 추출하였다. 또한 많은 검색엔진이 HEAD 부분의 meta file을 참조한 서비스를 하고 있기 때문에 HEAD 부분 태그를 제외시켰다. BODY 부분에서는 변수 명을 주는 <dfn>, <dt>, <dl>, 세 태그와 첨자체 <sup>, 보통 본문의 크기를 나타내는 <base font> 그리고 <table> 에 관련된 태그를 제외시켰다. 특히 table에 관련된 태그는 현재 웹페이지 경향 상 프레임이나 table을 가지고 디자인하기 때문에 table 에 관련된 태그는 제외시켰다.

가중치를 줄 수 없는 태그를 분리시키고 <a>, <marquee behavior>, , <blink>, , <b>, <big>, <cite>, , <u>, <var>, <i>, <ol>, <ul>, <strike>, <tt>, <h6>~<h1>, <font size> 18 종류의 태그를 추출하였다.

3.2 태그정보를 이용한 가중치 부여

추출된 태그에 다음과 같은 가중치를 부여하였다. 이에 대한 가중치 부여 순서는 웹상에서 300여건 이상의 임의의 문서에 대한 HTML 소스를 분석한 결과 아래 표1 과 같은 순위에 의해 중요한 색인어를 나타낸다는 것을 발견하였다.

<표 1> 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	

가중치 값은 현재에 웹상에서 가장 많이 본문으로 쓰고 있는 10포인트 글씨크기를 최하 기준가중치 값으로 하고 그 이상의 효과를 나타내는 태그에만 가중치를 주는 방향으로 하였다. 가중치의 범위는 1~9로 하였다. 10포인트 글씨체는 font size =3 과 같고 <h5> 보다 약간 큰 글씨다.

가중치 값은 가중치의 기준을 font size = 1~3 과 h5, h6 태그를 1점으로 하여 링크가 된 곳을 최고의 가중치를 9점을 주었다. 다음에 제목, 주제, 효과, 강조로 나누어 2~8까지의 가중치를 주었다.

3.3 가중치 테이블

웹문서 상에서 태그들은 하나만 쓰이는 것이 아니라 대부분이 2 개 이상이 중복해서 쓰이고 있다. 따라서 중복에 따른 가중치 테이블을 만드는 것은 꼭 필요하다.

그러나 태그가 2 개 이상 중복이 될 때도 웹상에서 문서 색인어의 가중치는 실질적으로 거의 상위 태그에 대한 가중치를 둘 수 있는 효과에 있기 때문에 어떠한 색인어에 태그가 3개 이상 나왔다 할지라도 상위 가중치를 가진 태그 2 개 만 가중치를 두는 것이 효과적이었다. 가중치 테이블은 <표 2>와 같이 가중치 분류목록을 기준으로 만들어 졌다.

가중치 값 분류기준 순서는 최고의 값, 링크값, 제목효과1, 제목효과2, 제목, 강조효과, 효과, 강조1, 강조2 그리고 항상 본문은 1로 가중치 값을 매겼다. 가중치 값은 최하 값 1에서 최고 값 10까지로 기준을 잡았다.

<표 2> 가중치 테이블 기준

분류	가중치	태그내용
최고어	10	<a> + span
링크	9	<a> + (을 제외한 태그)
제목효과1	8	+ (<a>을 제외한 태그)
제목효과2	7	h1,h2, font size=6 이상 + blink, marquee
제목	6	h1,h2, font size=6 이상 + 그 이하 point 태그
강조효과	5	태그 point 2, 3 글크기+ blink, marquee
효과	4	blink, marquee
강조1	3	태그 point 3 + 태그 point 2
강조2	2	태그 point 2 + 태그 point 1, 2
본문	1	본문

<표 3> 가중치 테이블의 일부

태그	...	span	h1~2	font size6	mar-quee	...
b	...	8	6	6	5	...
big	...	8	6	6	5	...
strong	...	8	6	6	5	...
u	...	8	6	6	5	...
strike	...	8	6	6	5	...

4. 가중치를 가진 색인어 추출 실험

HTML 문서를 받아들여 전처리 과정을 거친 후 형태소 분석기를 이용해 색인어를 추출한 다음, 추출된 색인어에 태그에 부여된 가중치 값을 부여하였다. 그리고 색인어 가중치 부여 가공작업 후에는 텍스트 파일 형태로 저장하거나 화면에 출력된다. 이를 그림으로 나타내면 그림 1 과 같다.

그림 1 에서 원정보는 텍스트화 되어있는 HTML 문서가 된다. 전처리과정에서 원정보에서 먼저 HEAD 부분을 자르고 원정보에 대한 한 문장씩 태그를 분석하고 가중치를 주는 태그가 발생하면 원정보를 형태소 분석기로 보내 유효한 정보 즉 색인어를 추출한다. 추출된 색인어를 분석된 태그 가중치에 의하여 색인어에 가중치를 부여하는 후처리과정을 거친다.

그리고 원정보의 수집 및 공급은 정보검색시스템 상에서는 웹에이전트가 담당하고 있으나 여기서는 본 연구에서는 수동공급으로 제한하고 있다.

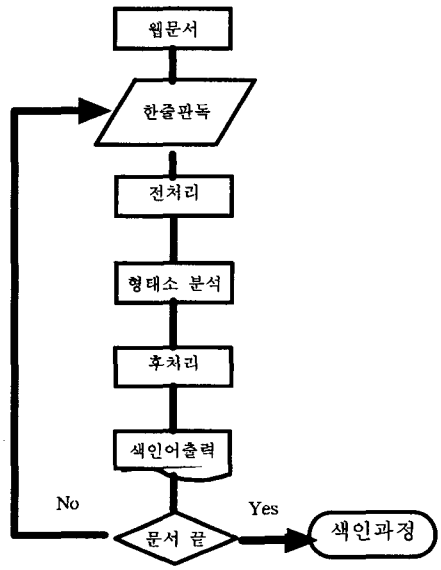

그림1 시스템 구조

4.1 색인실험

아래 HTML 소스는 실제 원문에서 <HEAD> 부분, 왼쪽 프레임 부분, 상단의 프레임 부분을 잘라낸 주 프레임의 일부분이다.

< HTML 문서>

.....

<B>1장. ATM 이란

1.1 지금까지의 교환 방식의 단점.

회선 교환 방식의 특징 (예: 전화망)

가장 고전적인 회선 교환 방식은 회선 접속 후 정보량이 많든 적든 간에 고정적인 속도에 의해 전송하므로 연속성이나 실시간

이 요구되는 음성통신에 알맞은 방식이나 한번 점유된 노드의 전송이 끝나기 전에는 다른 노드의 접속이 불가능한 단점 즉, 회선 효율성이 떨어지는 문제가 제기된다.

패켓 교환 방식의 특징 (예 : X.25)

패켓 교환 방식은 회선의 효율성을 강조 하기 위해 회선이 비기를 기다리지 않고 정보를 적당한 데이터 크기로 잘라 헤더에 수신처 주소, 제어정보를 추가해서 전송한다.

.....

위의 문서를 본 시스템에서 실험을 한 결과로 아래와 같이 나타났다.

<실행결과>

ATM@6	교환@3
교환방식@3	단점@3
회선@3	회선교환방식@3
패켓@3	패켓교환방식@3

결과에 의하면 가장 가중치가 높은 색인어로*ATM*을 추출하였고 다음 색인어로 *교환, 교환방식, 패킷교환방식, 단점, 회선, 회선교환방식, 패켓, 패켓교환방식*을 추출하였다.

그림 2 는 위 HTML 문서가 브라우저 상에서 실제 모습이다.

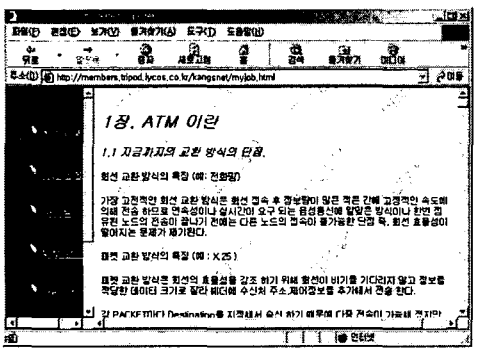

그림 2 실제 브라우저 상의 웹 문서

5. 결 론

정보검색시스템에서 검색된 문헌수가 많을 경우 문헌 순위화가 문헌 시스템의 성능을 좌우하는 요인이 된다.

본 논문에서는 검색된 문헌의 순위화 과정을 보다 효율적으로 수행하기 위하여 웹문서들은 태그로 이루어지며 중요한 색인어는 특정 태그 속에 포함되어 있다는 것에 근거하여 가중치에 영향을 줄 수 있는 태그를 추출하였다. 그 태그가 포함하고 있는 문자열에서 추출된 색인어에 그 태그 가중치를 부여함으로서 가중치를 가진 색인어를 추출할 수 있었다. 따라서 문헌 검색시 검색문헌이 가지고 있는 색인어 가중치 정보를 이용하여 문헌 순위화를 수행할 수 있다.

태그 정보를 이용한 가중치를 가진 색인기법은 색인어들이 다양한 가중치를 가지므로 기존의 빈도수 중심의 문헌순위화 방법에 비해 문헌순위화를 보다 효율적으로 수행할 수 있을 것이다.

그러나 같은 문서라 하더라도 HTML 문서를 작성하는 사람에 따라 다른 태그를 이용할 수 있으므로 보다 다양한 문서를 대상으로 하여 보다 표준화된 가중치 정보를 산출해내어야, 스타일시트를 활용한 문서 및 XML 문서 등에 대한 연구가 있어야 한다.

참고문헌

[1] McGill., M., M. Kill, and T. Noreault. " An Evalation of factors affecting document ranking by Information retrieval systems", Report from the School of Information Science , Syracuse University, Syracuse, New York, 1979.

[2] 양건묵, 박건일, 김유성, "한글 학술 논문의 일반 구조를 이용한 자동 색인어 선정 시스템", http://multibase1.inha.ac.kr/~kissess/work/index/index.htm

[3] 고미영, P-NORM 검색의 문헌 순위화 기법에 관한 실험적 연구, 연세대학교 대학원 박사논문, 1998.

[4] 정영미, 정보검색론, 구미무역, 1993.