한국산학기술학회논문지
Vol. 11, No. 9 pp. 3458-3464, 2010

사용자 중심의 블로그 정보 검색 기법

김승종[1*]
[1]한양여자대학 컴퓨터정보과

User-Centered Information Retrieving Method in Blogs

Seung-Jong Kim[1*]
[1]**Department of Computer Information, Hanyang Women's University**

요 약 최근 빠른 주기로 많은 양의 새로운 정보가 생성되기 때문에, 사용자 중심의 정보 검색을 위해 RSS라는 신디케이션 기술이 제공되고 있다. RSS는 새롭게 갱신된 콘텐츠를 자동으로 전달받을 수 있어 신규 정보를 찾기 위해 사이트에 지속적으로 접근하지 않아도 된다. 본 논문에서는 블로그 정보 검색을 위해 RSS 문서의 주소를 수집하는 수집기와 사용자 질의에 따른 RSS 문서의 순위결정 방법을 제안한다. 제안하는 정보 검색 기법을 이용하면 사용자가 RSS 문서를 효과적으로 검색할 수 있다.

Abstract Due to the recent tremendous growth of internet information, RSS, syndication technology provides internet users with a user-friendly information search. RSS enables you to automatically receive newly updated contents, so users do not need to constantly access web sites to obtain new information. This paper proposes the way of managing the web crawler, which collects the sites of RSS documents and helps the users efficiently use the RSS documents. And it also suggests the proper way of ranking the RSS documents based on the users' popularity. Users can efficiently search out the documents they need by using the proposed information searching methods.

Key Words : RSS, Document Recommendation, Document Rank

1. 서론

요즘은 빠른 주기로 많은 양의 새로운 정보가 생성되기 때문에, 사용자들이 필요로 하는 정보를 얻기 위해서는 다양한 검색과 웹사이트 서핑을 통해 정보의 유무를 확인해야 하고, 더욱이 원하는 페이지에 도달하기 위한 회원가입, 로그인, 검색 등 많은 과정을 거쳐야 하는 불편함이 있다. 이러한 불편함을 해소하기 위해 RSS(Really Simple Syndication)를 활용하는 방법이 제시되었으며, RSS는 웹 사이트를 통한 출판 과정에서 지속적으로 이루어지는 콘텐츠의 변화를 사용자들에게 자동 홍보하는 기법이다[1,2]. 하지만, RSS 기법을 이용하여 검색을 하더라도 검색 결과의 양이 매우 많아서, 사용자 자신에게 적합한 결과를 찾기 위해 또다시 노력을 기울여야만 한다. 그 이유는 대부분의 검색 기법이 사용자가 입력한 질의와의 일치도를 기준으로 검색을 수행하기 때문에 사용자가 입력한 질의어가 비적합 자료의 키워드와 일치하면 그 자료까지 결과로 제공되기 때문이다[3]. 따라서 정보의 가치가 증대됨에 따라 사용자의 관심과 선호도를 파악하여 보다 만족스러운 결과를 제공해주는 사용자 중심의 정보검색 기법의 필요성이 증대되고 있다[4,5]. 사용자 중심이라는 것은 사용자의 선택 및 수정에 기반을 둔 블로그 서비스를 의미한다. 기존의 정형화된 정보 검색 기법을 이용한 블로그 서비스는 사용자가 수동적으로 정보를 제공받거나 많은 시간을 들여서 정보를 검색해야 한다. 하지만 사용자 중심의 서비스는 사용자가 원하는 정보를 빠르고 집중적으로 제공받을 수 있도록 한다. 이러한 사용자의 요구를 만족시키기 위하여 문서 순위결정

본 논문은 2009학년도 한양여자대학 교내연구비지원에 의해 수행되었음.
*교신저자 : 김승종(jkim@hywoman.ac.kr)
접수일 10년 06월 28일 수정일 (1차 10년 07월 30일, 2차 10년 08월 11일) 게재확정일 10년 09월 08일

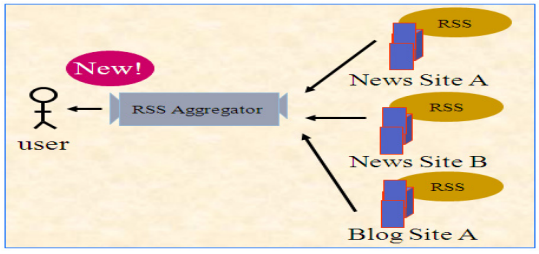
(Document Ranking) 또는 정보 필터링(Information Filtering) 방법에 관한 연구가 활발히 진행되고 있다. 하지만, 색인어와 질의에 따라 순위를 결정하면 사용자의 요구를 만족시키는 측면에서는 미흡한 점들이 많다[6,7].

본 논문에서는 사용자 스스로 관심 분야를 설정하고, 관심 분야와 관련된 초기 문서를 입력해 줌으로써, 사용자가 설정한 분야에 따라 분류되어 전달되는 RSS 리더와 각 분야를 대표하는 적합한 색인어를 추출할 수 있는 색인어 구성방법을 제시한다. 또한 사용자 프로파일(User Profile)을 구축하여 사용자의 선호도를 반영하고, 사용자의 요구에 적합한 문서를 적합성의 정도에 따라 제공하는 사용자 중심의 문서순위결정 기법을 제안한다.

2. 관련 연구

2.1 RSS

RSS(Really Simple Syndication)란 XML 기반의 간단한 콘텐츠 배급 프로토콜로서 블로그나 뉴스, 기업정보 사이트에 새로운 콘텐츠가 등록되었을 경우, 사용자들이 해당 사이트의 새로운 콘텐츠를 쉽게 전달받을 수 있도록 만들어진 포맷이며, RSS의 동작원리는 그림 1과 같다. 또한 새로운 콘텐츠를 전달받는 도구를 RSS 리더(reader)라고 하는데, RSS 리더는 블로그 뿐만 아니라 RSS 피드(feed)를 제공하는 모든 웹 사이트의 콘텐츠 정보와 이메일로도 정보를 가져올 수 있다[8,9].



[그림 1] RSS의 동작 원리

Kathleen Gilroy[10]는 RSS가 비즈니스와 정부 기관의 응용 등 사회 전반적인 분야에 활용할 수 있음을 보였으며, 작업 효율을 높여서 생산성을 향상시키고 적지적소에 필요한 정보를 제공할 수 있음을 RSS 구현사례를 통해 입증하였다. Huang[11]은 지능적인 시스템에서 문맥의 의미를 부여해야 할 때, RSS 기술을 이용하여 지식 구조를 모델링 하고, 메타 정보를 제안하여 사용자와 시스템 간의 상호작용을 높일 수 있음을 보여주었다. 위에서 언급한 RSS 관련 연구는 사용자가 이용하는 웹페이지를 구분하기 위하여 특징정보를 통해 문서를 분류하고, 이를 통하여 사용자들의 웹 내비게이션 패턴을 인식하고 해당 사용자에게 콘텐츠를 추천하는데 의미를 두고 있다.

2.2 정보 필터링과 문서 순위 결정

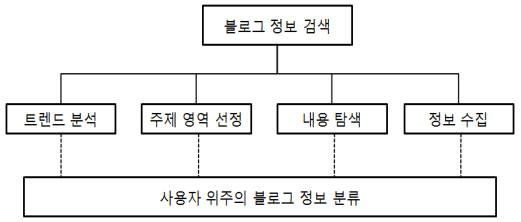
정보 필터링 방법은 사용자의 선호도를 저장한 후, 선호도 판단에 필요 없는 정보를 제거하여 검색된 결과의 수를 줄이는 방법이다. 그러나 사용자의 관심이 변하지 않거나 정보자원에 커다란 변동이 없으면 계속 사용할 수 있으나, 특정 관심분야로 한정된다는 단점이 존재한다[7,12,13]. 정보 필터링 시스템에서 정보 요구의 표현을 프로파일(profile)이라고 하며, 프로파일이 정보 검색과 데이터베이스 시스템에서 사용되는 질의와 같은 역할을 하므로 정보 필터링 분야의 핵심이 되고 있다. 이러한 프로파일 모음을 사용자 모델이라고 한다. 현재 이용되고 있는 필터링 기법은 크게 규칙기반 필터링, 내용기반 필터링, 그리고 협조적 필터링으로 구분된다[14]. 규칙기반 필터링 기법은 미리 정의된 규칙에 맞는 적합한 정보만을 걸러서 간단히 제시하는 시스템에 적용된다. 내용기반 필터링 기법은 사용자 프로파일과 정보 간의 유사성을 고려하여 필터링하는 방법으로서, 가장 많이 사용하는 기법이다. 협조적 필터링은 유사한 관심분야의 다른 사용자의 학습된 선호도를 참고하여 사용자 피드백을 통해 확장 또는 축소를 수행하여 개인의 선호도를 지속적으로 구축해 나가는 방식이다[12, 14]. 인터넷 상에서 서비스하는 대부분의 정보 필터링 시스템은 사용자 프로파일 구축을 위해 키워드 벡터와 적합성 피드백 방법을 결합하여 사용한다[15].

문서순위(Document Rank)결정은 사용자의 질의어와 검색된 문서들이 얼마나 유사한가에 따라 문서의 우선순위를 결정하고, 이 순서에 따라 사용자가 가장 적합한 문서를 참조하도록 하는 방법이다. 문서순위결정 기법에는 가중치와 유사도, 적합성 피드백, 데이터 융합, 불리언(Boolean) 모델 등이 있다. 이 중에서 가장 많이 사용하는 문서순위결정 기법은 가중치와 유사도를 이용하는 방법으로, 전체 문서에서 추출한 색인어 벡터로 문서와 질의를 표현하고, 색인어의 가중치와 유사도를 계산하여 유사도가 큰 순서대로 문서의 순위를 결정하는 방법이다[14-16].

3. 제안하는 블로그 정보 검색 기법

3.1 블로그 정보 검색의 개념

블로그 정보 검색은 그림 2와 같이 먼저 검색 트렌드(trend)와 주제 영역을 선정한 후, 내용을 탐색하면서 정보를 수집하고 적합한 결과를 사용자에게 분류하여 보여주는 과정으로 진행된다.



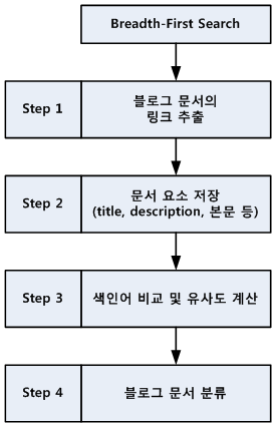
[그림 2] 블로그 정보 검색의 개념도

3.2 블로그 정보 검색 과정

블로그 정보 검색에서 블로그 스파이더(Spider)는 일반적으로 Random Search 알고리즘을 사용하지만, 본 논문에서는 Breadth-First Search 알고리즘을 사용한다. 왜냐하면 Random Search 알고리즘은 크롤러(Crawler)의 움직임을 제어할 수 없고 또한 임의의 주제에 대한 정보를 수집하는 반면, Breadth-First Search 알고리즘은 목표를 찾을 때까지 특정 주제에 대해 집중적으로 검색하거나 수집할 수 있는 장점이 있다. 그림 3은 제안하는 블로그 정보 검색 알고리즘의 각 단계별 블록도를 나타낸 것이며, 자세한 내용은 다음과 같다.

Step 1 : 블로그 문서의 링크 추출

블로그 정보 수집기는 메인 페이지에 존재하는 링크의 수로 탐색 큐의 크기를 조절한 후, Breadth -First Search 알고리즘을 이용하여 블로그 문서에서 링크들을 추출한다.



[그림 3] 제안하는 블로그 정보 검색 알고리즘

Step 2 : 문서 요소 저장

블로그 메인 페이지가 프레임 형태로 되어 있을 경우, 각 프레임에 해당하는 주소를 획득하여 실제 블로그 메인 페이지 주소를 재설정한다. 재설정된 블로그의 메인 페이지에 존재하는 링크주소를 획득하기 위해 표 1과 같은 정규표현식(Regular Expression)을 사용한다[2,8,9]. 수집기는 획득된 링크주소들에 접근하여 블로그 채널이 가져야 할 태그 여부를 확인한 후, 블로그 채널이라고 판단되면, 이미 탐색된 채널인지, 아닌지를 비교하고 신규 채널일 경우 해당 블로그 채널의 주소와 문서의 요소(title, description, 본문 등)를 저장한다.

[표 1] RSS 탐색을 위해 사용된 정규 표현식

패턴 종류	정규 표현식
프레임 태그	<frame\\s+(.*?)\\s+src\\s*=\\s*\"?(.*?)[\"|>]
링크 태그	<a\\s+href\\s*=\\s*\"?(.*?)[\"|>]
RSS 문서 태그	^<channel

Step 3 : 색인어 비교 및 유사도 계산

RSS 문서로부터 획득한 링크주소를 이용하여 가져온 실제 문서는 분류모듈에서 어절 단위 분리, 조사, 불용어 제거 작업을 수행하고 문서 내 단어의 빈도수를 기록한다. 추출된 모든 단어와 단어빈도수는 웹 콘텐츠 분류 시에 사용된다. 색인어 추출은 대형 포털 사이트의 트렌드와 주제영역을 분석한 후, 그 주제 영역별 100개의 문서를 대상으로 실험한 결과 문서 당 2회 이상 추출된 단어를 색인어로 추출하였고, 이렇게 추출된 대표 색인어 집단은 표 2와 같다. 입력된 문서 내에서 추출된 단어들은 식(1)과 같이 많이 알려진 유사도(Similarity) 계산 방법을 이용하였으며, 분야별 색인어 집단과의 유사도를 계산하여 가장 큰 값을 갖는 분야로 분류한다[16,17].

[표 2] 주제영역과 대표 색인어 추출 결과

주제 영역	대표 색인어 추출 결과
IT	컴퓨터,하드웨어,프로그램,프로그래밍,소프트웨어,웹,인터넷,IT,운영체제,윈도우,유닉스,리눅스,보안,통신,미디어,3G,애플,구글,2.0,모바일,네트워크,폰
예술	그림,공연,연극,무용,미술,디자인,예술,문화,갤러리,로모
경제	주식,경제,부동산,기업,증권,재테크,기업,산업,유통,코스닥,투자,환율,지수,청약,임대,금리,은행,펀드,대출,보험,거래,나스닥,감정평가,금융

교육	등록금,교육,영어,수학,논술,과학,입시,이공계,english,초등,중등,고등,학교,공부,시험,대학,임용,수능,학생,교사,교수,사립
연예	연예,드라마,소녀시대,무한도전,영화제,유재석,패션,팬미팅,열애,탤런트,연예인,배우,모델
사건	사건,살인,절도,폭행,불륜,범인,자살,화재,교통,검거,재난,사고
게임	게임,맞고,고스톱,바둑,장기,halo,마비노기,아이템,위닝,PSP,닌텐도,온라인게임,스타크래프트,리니지,카트라이더,프리스타일
여행	휴가,여행,여행지,관광지,신혼,해외여행,해수욕장,렌터카,배낭,골프,허니문,펜션,항공,호텔,관광,유럽,여권
스포츠	스포츠,농구,축구,야구,선수,우승,k-1,수영,골프,탁구,MLB,메이저리그,맨유,FA,리그
만화	스쿨데이즈,무협,순정,만화,애니,그렌라간,드래곤볼,코난
문화	소설,문학,수필,도서,독자,도쿄타워,리뷰,도서평,문고,영화,movie,관객,극장,흥행,예매,시사회,음악,ost,발라드,클래식,재즈,rock,오페라,music,pop,팝,힙합,트로트
정치	정치,선거,경선,한나라당,민주당,외신,회담,대통령,국회,회담,박근혜,노무현,이명박,오세훈,한명숙,외교,북한,국방,청와대,의원,정부

$$S(d, q) = \sum_{i=1}^{n} (w_i^d \times w_i^q) \quad (1)$$

식(1)에서 d는 블로그 문서, q는 질의어를 각각 의미한다. 또한 w_i^d는 문서 d에서 i번째 색인어의 가중치, w_i^q는 질의어 q가 i번째 색인어에서의 가중치를 각각 의미하며, 식(2), (3)과 같이 정의된다.

$$w_i^d = (\log(f_i^d) + 1.0) / \sum_{i=1}^{n} [\log(f_i^d) + 1.0]^2 \quad (2)$$

$$w_i^q = \frac{(\log(f_i^q) + 1.0) \times \log(N/n_i)}{\sum_{i=1}^{n} [(\log(f_i^q) + 1.0) \times \log(N/n_i)]^2} \quad (3)$$

식(2) 및 (3)에서 f_i^d는 문서 d에서 i번째 색인어의 출현 빈도, f_i^q는 질의어 q가 i번째 색인어에서의 출현 빈도를 각각 의미하고 N은 전체 블로그 문서의 개수, n_i는 i번째 색인어가 포함된 블로그 문서의 개수를 의미한다.

Step 4 : 블로그 문서 분류

색인어의 빈도수와 색인어가 나타나는 문헌의 빈도수를 조사하고 분야별 색인어 집단을 선정한 후, 고려해야 할 것은 다른 분야와의 색인어 중복이다. 본 논문에서는 타 분야의 색인어로 선정된 것은 해당 분야에서 제거시키는 방법을 제안하여 분야별 블로그 문서 분류 시, 효율성을 가지는 색인어 집단을 구성하도록 한다. 색인어를 토대로 중복 문서를 분류하는 기준은 식(4)와 같다.

$$S_2 > S_1 \times 0.9 \quad (4)$$

식(4)에서 S_1은 1순위 유사도 값, S_2는 2순위 유사도 값을 각각 의미하며, 식(1)로부터 계산된 값이다. 가중치를 0.9로 설정한 이유는 표 3과 같이 분야별로 임의로 선정한 20개의 문서에 대해 실험한 결과, 분류된 결과가 다르면서 S_2/S_1가 0.9를 초과하는 경우에는 다른 문서로 분류되었기 때문이다. 즉, 식(4)의 조건을 만족하면 다른 문서로 분류되고 주제 영역도 정상적으로 분류되며, 만족하지 않으면 해당 블로그 문서는 제거된다.

3.3 블로그 문서의 순위 결정

RSS는 콘텐츠를 자동으로 전달하기 때문에 블로그 검색 도구는 주로 블로그의 제목과 소개 등을 대상으로 검색이 이루어지므로 정보 갱신주기와는 무관한 순위가 제시된다. 따라서 블로그 정보가 갱신되는 문서들을 그대로 제시해 주기보다는 우선순위를 정하여 제시해줄 필요성이 있다.

[표 3] 임의 문서 20개의 유사도 측정 결과

No	입력문서의 단어개수	추출된 색인어 개수	블로그 게시자가 분류한 분야	제안한 방법 (1순위)	제안한 방법 (2순위)	S2/S1	비고
1	603	341	국제	국제	문화	1.81	
2	225	136	정치	정치	문화	1.10	
3	332	206	스포츠	스포츠	문화	1.48	
4	138	99	경제	경제	정치	5.44	
5	295	208	사회	사회	사회	1.81	
6	133	84	IT	IT	사회	1.63	
7	508	326	문화	문화	경제	2.71	
8	86	64	사회	사회	문화	2.00	
9	508	326	국제	문화	정치	0.67	제거
10	254	157	정치	정치	사회	1.31	
11	85	62	문화	문화	사회	3.33	
12	362	277	경제	국제	경제	1.23	
13	209	129	정치	경제	국제	0.72	제거
14	207	147	경제	사회	경제	0.47	제거
15	221	153	스포츠	스포츠	경제	1.90	
16	125	78	IT	IT	문화	2.35	
17	203	109	IT	IT	사회	0.98	
18	186	132	스포츠	스포츠	국제	1.19	
19	267	176	스포츠	스포츠	문화	3.50	
20	85	42	문화	문화	국제	3.10	

Yuwono[17]은 문서 단위의 유사도를 계산하여 문서의 순위를 결정하는 방법을 제안했지만, 사이트 단위의 우선순위를 결정하는 데에는 무리가 있다. 즉, 사용자가 질의한 단어를 다수 포함한 글이 1개 등록된 RSS 채널과 사용자가 질의한 단어 1개만을 포함한 다수의 글이 등록된 RSS 채널을 비교해 보자. 단순한 유사도 계산만으로 순위를 결정하면 전자가 높은 우선순위를 가질 수 있는 단점이 존재한다. Brin[18] 등은 정적인 링크를 대상으로 블로그 문서의 순위를 결정하는 방식을 제안했으나, 타 블로그와의 하이퍼링크 정보는 유동적으로 변하며, 링크의 수를 제한해 놓은 경우가 대부분이기 때문에 블로그 문서의 순위가 잘못 결정될 가능성이 많다. Agarwal[3] 등은 다른 블로그 문서 등과 연결된 링크의 수와 댓글의 수로 블로그 문서의 순위를 결정하여 영향력을 판단하는 모델을 제안했으나, social network 관점에서 인기도를 파악하긴 좋지만, 다른 블로그 문서를 인용하는 링크의 수는 무분별한 복사와 인용이 발생하므로 부적절한 기준이 될 가능성이 많다. 또한 블로그 문서에 관심 있는 사용자들에 의한 댓글 정보인지를 판단할 수 있는 근거가 부족하기 때문에 의미 없는 정보들에 의해 블로그 문서의 순위가 잘못 결정될 가능성이 많다.

본 논문에서는 검색 기간 내에 블로그의 갱신주기가 짧으면서 내용이 꾸준히 갱신되는 문서에 대해 높은 우선순위를 부여하는 알고리즘을 제안한다. 즉, 식(5)를 적용하여 1차적으로 DR_1 값이 큰 블로그 문서에 대해 높은 우선순위를 부여하고 만약, 식(5)를 적용한 결과가 동일하면, 식(8)과 같이 블로그 갱신주기를 반영하여 DR_2 값이 작은 블로그 문서를 사용자에게 제공한다.

$$DR_1 = DR_{title} + DR_{text} \tag{5}$$

$$DR_{title} = q / \sum_{i=1}^{n} (m_i \times 1.0) \tag{6}$$

$$DR_{text} = q / \sum_{i=1}^{n} (m_i \times 0.5) \tag{7}$$

식(5)에서 DR_1은 검색기간 내, 블로그 문서의 순위 값을 의미하며, DR_{title}은 블로그 문서의 제목(title) 순위 값, DR_{text}는 블로그 문서의 본문(text) 순위 값을 각각 의미한다. 또한 DR_{title}과 DR_{text}의 가중치를 각각 1.0과 0.5로 설정한 이유는 질의어가 제목에 위치하면 상대적으로 본문에 위치한 경우보다 높은 우선순위를 부여하기 위함이다. 식 (6)에서 n은 하루 단위의 검색기간을 의미하며, m_i는 i번째 검색 일에 질의어 q와 일치하는 문서의 개수를 의미한다. 참고로 사용자 질의어 q는 질의되는 단어의 개수를 의미한다. 식(5)는 블로그 검색 기간 내, 질의어를 포함하는 블로그 문서의 제목과 본문의 개수에 따라 순위를 결정하게 된다.

예를 들어, 블로그 검색 기간을 10일로 설정하고 사용자 질의어 q는 1로 가정하자. 또한 제목 부분에만 질의어 q가 포함된 블로그 A와 블로그 B가 있다고 가정하자. 검색 기간에 질의어 q를 포함한 문서의 개수가 블로그 A는 {1, 2, 2, 2, 4, 2, 2, 3, 1, 1}, 블로그 B는 {0, 0, 0, 5, 4, 5, 0, 6, 0, 0}라고 조사되었을 때, 식(6)에 의해 DR_{title}을 각각 계산하면 다음과 같다.

블로그 A: DR_{title}=1/(1+2+2+2+4+2+2+3+1+1)=1/20

블로그 B: DR_{title}=1/(0+0+0+5+4+5+0+6+0+0)=1/20

검색 기간 내에 블로그 A 및 B의 DR_1은 동일하므로 식(8)과 같이 블로그의 갱신주기를 반영하여 DR_2 값이 작은 블로그 문서, 즉 자주 갱신되는 문서가 우선적으로 사용자에게 제시되도록 한다.

$$DR_2 = \frac{1}{T} \sum_{j=1}^{T} (x_j - M)^2 \tag{8}$$

식(8)에서 T는 블로그 검색 기간을 블로그 갱신 주기로 나눈 값이며, x_j는 j번째 블로그 갱신주기에 질의어가 포함된 문서의 개수를 의미한다. 또한 M은 블로그 검색 기간 내에 평균적으로 갱신된 문서의 개수를 의미한다. 위의 예에서 갱신주기를 2일로 설정하면 블로그 A의 x={3, 4, 6, 5, 2}, 블로그 B의 x={0, 5, 9, 6, 0}이다. 두 블로그의 평균은 4로 동일하며, 식(8)을 이용하여 DR_2를 계산하면 다음과 같다.

블로그 A : DR_2 = (1+0+4+1+4)/5 = 2

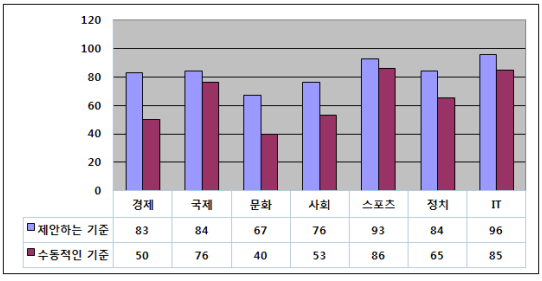
블로그 B : DR_2 = (16+1+25+4+16)/5 = 12.4

검색 기간 내 질의어를 포함한 문서의 수는 같지만, DR_2를 계산해 보면 블로그 A가 블로그 B보다 작은 값을 가지므로 우선순위가 높다. 본 논문에서 제안한 블로그 문서의 순위결정 알고리즘은 지정된 검색 기간 내에 문서가 얼마나 자주 갱신되는지의 여부를 체크하여 사용자에게 효율적인 블로그 문서의 제공이 가능할 것이다.

4. 실험 결과 및 고찰

4.1 색인어 추출에 따른 문서 분류 결과

본 논문에서 제안한 블로그 문서 분류 기법의 우수성을 입증하기 위해 색인어 집단 추출 시 사용되었던 분야당 100개의 블로그를 이용하였다. 또한 5명의 블로그 게시자에 대해 분야별 최저 5개에서 최대 20개의 색인어를 선정하도록 하였으며, 개인별로 선정한 색인어 중에서 중복된 색인어를 제거하였다. 실험결과는 그림 4와 같으며, 블로그 게시자가 직접 색인어를 추출하여 문서를 분류하는 방식(수동적인 기준)과 자동화된 기준으로 색인어를 추출하고 이를 토대로 문서를 분류하는 방식(제안하는 기준)에 대해 각각 비교하였다. 그림에서도 알 수 있듯이 제안하는 방식을 이용하여 문서를 분류한 결과 평균 83.3%, 수동적인 기준에 의한 결과는 평균 65.0%의 정확성을 나타내었다. 모든 분야에서 제안하는 방식이 정확하게 분류되었으며, 특히 다양한 분야가 섞여 있는 경제, 문화, 사회에서는 수동적인 기준에 의한 방법과 비교하여 약 20% 이상 차이가 날 정도로 색인어 집단이 잘 구성되었다고 볼 수 있다.



[그림 4] 색인어 추출에 따른 문서 분류 결과

4.2 문서 검색 결과

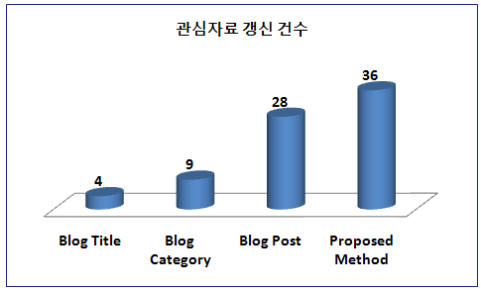
기존의 메타블로그 사이트(블로그 코리아, AllBlog, POPCON BLOZINE 등)에서 제공하는 블로그 검색 도구는 블로그 제목 및 소개 검색, 블로그 종류 검색, 블로그 신규 게시물 검색 등을 제공한다. 표 4는 대표적인 메타블로그 사이트에서 제공하는 블로그 문서 검색 방식과 제안하는 방식을 비교한 것이다. 메타블로그 사이트와 비교했을 때, 기능은 유사하지만, 입력요구사항이 적고 문서의 자동 수집 기능이 있으며, 가장 큰 장점은 블로그 문서의 분류 체계를 유동적으로 구성할 수 있다는 것이다. 즉, 기존의 메타블로그 사이트에서는 블로그 문서를 사용자가 수동으로 분류하기 때문에 만약 잘못된 분류일지라도 문서의 카테고리를 수정하기가 쉽지 않다. 하지만 제안한 방식은 사용자가 수동으로 분류한 블로그 문서를 자동으로 재분류할 수 있으므로 사용자 중심의 검색서비스 제공이 가능하다.

블로그 검색 성능을 비교하기 위해 검색 기간은 45일로 설정하고 엠파스 포털사이트에 상위 랭크된 10개 블로그의 RSS 주소를 획득하였다. 엠파스에 저장된 상위 10개 블로그의 RSS 채널수는 8,750개였으며, 제안하는 방식을 이용하여 획득한 상위 10개 블로그의 RSS 채널수는 1,307개였다. 따라서 약 1/6.7 정도 줄어든 RSS 채널을 확보하여 정확성을 향상시켰으며, 최근 15일간 질의어 "여행"을 포함하는 신규 자료의 갱신 정도를 그림 5에 나타내었다. 그림에서 Blog Title은 블로그 제목을 대상으로 질의어 "여행"을 포함하는 갱신된 블로그 문서의 수를 의미하고 Blog Category는 소주제, 대주제로 분류한 Blog Post의 묶음을 대상으로 갱신된 블로그 문서의 수를 의미한다. 또한 Blog Post는 블로그 사이트에 등록된 description을 대상으로 질의어 "여행"을 포함하는 갱신된 블로그 문서의 수를 말한다. 그림 5에서 알 수 있듯이, 실제 엠파스 포털사이트에서 제공하는 검색 알고리즘(Blog Post) 보다 제안하는 방식이 자주 갱신되는 블로그 문서를 효율적으로 검색할 수 있었다.

일반적인 검색 사이트에서는 자료의 갱신정도를 검색 알고리즘에 포함시키지 않았지만, 제안하는 방식은 사용자 질의어를 토대로 갱신 주기 및 갱신 분포 등을 고려하여 블로그 문서를 검색하기 때문에 사용자의 요구를 반영한 최적의 블로그 문서를 제공할 수 있다는 장점이 있다.

[표 4] 메타블로그 사이트의 검색 방법 비교

	블로그 코리아 (www.blogkorea.org)	AllBlog (www.allblog.net)	POPCON BLOZINE (www.blozine.com)	제안하는 방법
수집 정보	블로그 주소 RSS 주소 블로그 제목 블로그 소개 블로그 종류	블로그 주소 블로그 소개	블로그 주소 RSS 주소 블로그 제목 블로그 소개 블로그 종류	블로그 주소 RSS 주소 블로그 제목 블로그 소개 블로그 종류
질의어 입력 방법	사용자 직접 입력	사용자 직접 입력	사용자 직접 입력	사용자 직접 입력
수집 방법	주기적인 자동 수집	주기적인 자동 수집	주기적인 자동 수집	주기적인 자동 수집
추가 기능	없음	없음	고정적인 블로그 분류	유동적인 블로그 분류



[그림 5] RSS 채널 검색의 성능 비교

5. 결론

다양한 정보와 블로그의 증가로 사용할 수 있는 RSS 채널은 많아졌지만, 사용자는 각 블로그를 직접 돌아다니며 내용을 판단하고 RSS 채널 주소를 찾아내야 하는 문제점이 있다.

본 논문에서는 사용자 중심의 효율적인 블로그 정보 검색을 위하여 문서의 색인어 추출 방법, 추출된 색인어를 이용한 문서 분류 방법과 순위 결정 방법을 제안하였다. 사용자는 관심 있는 분야를 설정하여 관련 문서를 제안 알고리즘에 넣어주면, 해당 분야의 색인어 집단이 자동으로 추출되고, RSS 채널을 통하여 들어오는 정보는 이를 기준으로 분류되어 사용자에게 제공된다. 또한 RSS 문서 자동 탐색 기능과 함께 사용자의 다양한 질의에 대하여 RSS 채널을 추천해 주기 때문에 사용자는 자신의 관심 분야에 대한 RSS 문서를 효율적으로 검색하고 제공받을 수 있다.

참고문헌

[1] KISTI, RSS를 이용한 웹페이지의 뉴스 피드 기능, 2005.

[2] World Wide Web Consortium, 2005.

[3] N. Agarwal and H. Liu, “Blogosphere: Research Issues, Tools, and Applications”, SIGKDD Explorations, 10(1): 18 - 31, July, 2008.

[4] K. C. Sia, J. Cho, C. Yun, B. L. Tseng, “Efficient Computation of Personal Aggregate Queries on Blogs”, Proc. Knowledge Discovery and Data Mining Conf., ACM Press, pp. 632-640, 2008.

[5] A. Stewart, L. Chen, R. Paiu, and W. Nejdl, “Discovering Information Diffusion Paths From Blogosphere for Online Advertising”, Proc. Workshop on Data Mining and Audience Intelligence for Advertising in conjunction with Knowledge Discovery and Data Mining, ACM Press, pp. 46-54, 2007.

[6] Bracha Shapira, et al., "Information Filtering: A New Two-Phase Model using Stereotypic User Profiling," Journal of Intelligent Information systems, Vol. 8, 1997.

[7] Czeslaw Danilowicz, Jaroslaw Balinski, "Document Ranking based upon Markov Chains", Information Processing and Management, Vol.37, pp. 623-637, 2001.

[8] Kathleen Gilroy, Winning the Race for Knowledge Worker Productivity, A White Paper prepared for the Int. Conference on the National Communications Commission, pp. 3-23, 2005.

[9] RSS Technology Reports, 2005.

[10] PEW INTERNET & AMERICAN LIFE PRO -JECT, 2004.

[11] Weihong Huang, “Enabling Context-Aware Agents to Understand Semantic Resources on the WWW and The Semantic Web”, Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 138-144. 2004.

[12] Douglas W. OARD, "The State of the Art in Text filtering," User Modeling and User-adapted Interaction, vol.7, pp. 141-178, 1997.

[13] Foltz, P. W, "Using Latent Semantic Indexing for Information Filtering," Proceedings of the Conference on Office Information Systems, Cambridge, MA, pp. 40-47, 1990.

[14] Passamo, M. and Billsus, D., "Learning and Revising User Profiles: the Identification of Interesting Web Sites", Machine Learning, Vol. 27, pp. 313-331, 1997.

[15] Dwi H. Widyantoro, Thomas R. loerger, John Yen, "An Adaptive algorithm for Learning Changes in User Interests," 8th International Conference on Information and Knowledge Management(CIKM'99), November 2-6, Kansas city, 1999.

[16] Michael Persin, "Document Filtering for Fast Ranking," ACM-SIGIR, pp. 339-348, 1994.

[17] B. Yuwono, "Search and ranking algorithms for locating resources on World Wide Web", Proc. of the Int. Conf. on Data Engineering, pp. 164-171, 1996.

[18] Brin, S. & Page, L., "The Anatomy of a Large-Scale Hyper-textual Web Search Engine", Computer Networks and ISDN Systems, pp. 1107-1117. 1998.

김 승 종(Seung-Jong Kim) **[정회원]**

- 1994년 2월 : 한양대학교 전자통신공학과 (공학석사)
- 2000년 8월 : 한양대학교 전자통신공학과 (공학박사)
- 2000년 9월 ~ 현재 : 한양여자대학 컴퓨터정보과 교수

<관심분야>
영상처리, 영상통신, 멀티미디어 응용