개인화된 학습서비스를 위한 관심분야에 따른 웹 문서 분류기

김준일[0], *이영석, **조정원, ***최병욱
[0]한양대학교 정보통신공학과
*한양대학교 전자통신컴퓨터공학과
**제주대학교 컴퓨터교육과
***한양대학교 정보통신학부
{jjunil, yslee38, bigcho, buchoi}@mlab.hanyang.ac.kr

Web Document Classifier based on Interesting Field for Personalized Learning Service

Jun-Il Kim[0], *Youngseok Lee, **Jungwon Cho, ***Byung-Uk Choi
[0]Department of Information Communication Engineering, Hanyang University,
*Department of Electrical and Computer Engineering, Hanyang University,
**Department of Computer Education, Cheju National University,
***Division of Information and Communication, Hanyang University

요 약

RSS와 같은 신디케이션 기술은 사용자가 스스로 웹사이트에 접근하지 않아도, 새롭게 업데이트 되는 정보가 있을 때마다 RSS Aggregator를 통해 사용자에게 알려줌으로써 편리성을 가져다준다. 이러한 기술을 이용한다면 학습자들은 새로운 웹 문서가 게시될 때마다 해당 웹사이트를 방문할 필요 없이, 자동으로 신규 정보만 얻어오는 학습 자료의 습득 도구로서 활용이 가능하다. 하지만, 정보가 새롭게 추가되는 여부만을 판단하는 기존의 RSS Aggregator의 경우에 등록된 채널수가 늘어갈수록 사용자는 자신이 원하는 정보를 찾기 위해, 정보를 분류하고 검색하는 작업에 많은 노력을 기울여야한다. 본 논문에서는 이러한 문제점을 해결하고 사용자에게 보다 효율적인 정보 전달을 위해, 사용자 스스로 관심분야를 생성하고, 사용자에게 전달되는 신규자료는 각 분야에 자동적으로 분류되며, 사용자가 지정해 놓은 조건에 적합하도록 콘텐츠를 제공 받을 수 있는 시스템을 설계하였다. 신규자료를 분야에 자동적으로 분류하기 위해 초기 분류된 문서로부터 분야별 색인어 추출 방안을 제안하고자 한다.

1. 서 론

우리가 찾고자 하는 정보의 대부분이 인터넷이라는 정보의 바다에 존재하지만, 원하는 정보를 찾는 일은 쉽지 않은 일이다. 빠른 주기로 많은 양의 새로운 정보가 생성되기 때문에, 사용자들이 이러한 정보를 얻기 위해서는 다양한 검색과 지속적으로 웹사이트를 돌아다니며, 신규 정보의 유무를 확인하는 불편을 감수해야한다.

또한, 사용자는 원하는 페이지에 도달하기 위해 회원가입, 로그인, 검색 등 많은 과정을 거쳐야 하기 때문에, 정보 수집에 불편함이 있다. 이러한 불편함을 해소하기 위해, 사용자에게 신규 정보를 전달하기 위한 대체 방안으로 이메일이 활용되었으나, 많은 스팸메일로 인해 그 활용도는 떨어지고 있는 추세이다.

현재는 이메일의 대안으로 RSS가 활용되고 있다. RSS는 현재 뉴스, 블로그 사이트를 기반으로, 최근 추가되거나 변경된 페이지들에 대한 요약 정보를 제공한다.

2004년 PEW ONTERNET 조사에 따르면 미국의 1억 2천만 성인 인터넷 사용자 중 7%가 자신의 블로그를 만든 경험이 있고(800만명), 27%가 블로그를 주기적으로 구독하고 있으며, 5%가 RSS 뉴스 Aggregator(Reader)를

사용하고 있다[1]. 사용자는 RSS Aggregator에 해당 RSS 파일의 위치만 기억해두면, 나중에는 이 파일에 요약된 내용만 보고 중간 과정 없이 직접 해당 페이지로 바로 접근할 수 있게 되므로 방문자들의 수고가 크게 줄어든다. 이는 교육자료 습득의 도구로서도 쓰일 수 있다. 교육 자료로 쓰일 수 있는 뉴스나 정보 사이트의 RSS 채널 주소를 등록하여 주기적으로 새로운 정보를 얻어옴으로써 교육 효과를 높일 수 있다[2,3].

하지만 현재의 RSS Aggregator는 다음과 같은 문제점을 지니고 있다. 첫째, RSS 문서 내부에 존재하는 실제 문서의 내용 정보가 극히 적어 사용자는 RSS 문서만을 보고 실제 내용을 제대로 판단해내기 어렵다. 둘째, RSS Aggregator에 등록된 채널이 많아질수록, 사용자에게 전달되는 정보의 양도 증가한다. 따라서 사용자는 자신이 원하는 학습 자료를 찾기 위해 또 다른 수고를 감수해야한다.

따라서 본 논문에서는 위 문제를 해결하여 개인화된 학습서비스가 가능하도록 하는 웹문서 전달 시스템을 제안한다. 본 시스템에서는 사용자 스스로 관심 분야를 설정하고, 여기에 초기 문서를 입력해 줌으로써, 차후에 RSS를 통하여 들어오는 신규 데이터들은 사용자의 설정한 각각의 분야대로 분류되어 전달된다. 이를 위해서 초기 문서에서 각 분야를 대표하는 적합한 색인어 추출을 할 수 있도록, 효율적 색인어 구성방안을 실험을 통해 제시한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구에 관해 기술한다. 3장에서는 본 논문에서 제시하는 개인화된 웹 콘텐츠 전달 시스템의 구조와 기능에 관해 기술한다. 4장에서는 분류를 위한 색인어 집단 선정의 분석과 실험결과를 살펴보고, 5장에서 결론 및 향후 과제를 기술한다.

2. 관련 연구

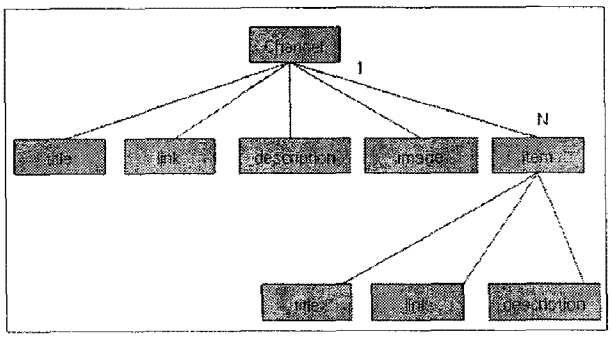
2.1 RSS

Really Simple Syndication, 혹은 Rich Site Summary의 줄임말이며, XML(eXtentional Markup Language) 혹은 RDF(Resource Description Framework) 기반의 콘텐츠 배급 프로토콜이다[4]. 이는 웹정보 제공자 측에서 새로운 정보의 갱신 여부를 알려주는 용도로 쓰인다. RSS는 RSS Aggregator라는 소프트웨어를 통해 사용자에게 보여 진다. 현재 RSS는 <표1>과 같이 7가지 버전이 존재한다[5].

<표 1> RSS 버전별 특징

버전	오너	설명	진행
0.90	넷스케이프	초기 버전	1.0에 의해 중단
0.91	유저랜드	0.90 간략화	2.0에 의해 중단되었지만, 많이 쓰임
0.92, 0.93, 0.94	유저랜드	0.91확장	2.0에 의해 중단
1.0	RSS개발 그룹	RDF기반	안정화코어, 모듈개발 진행 중
2.0	유저랜드	0.91확장	안정화코어, 모듈개발 진행 중

RSS는 각 버전별로 약간의 차이를 가지지만 대부분 <그림1>와 같은 구조로 구성되며, 각 요소는 다음과 같은 정보를 포함한다.



<그림 1> RSS 구조

-Title: 채널 이름.
-Link: 채널의 주소 (RSS 등록 주소).
-Description: 채널의 간단한 소개 정보.
-Image: 채널의 로고 이미지 포함 여부.
-Item: 채널에서 업데이트된 새로운 콘텐츠.
-Item.title: Item의 제목.
-Item.link: Item의 링크 주소.
-Item.description: 해당 Item의 요약 내용.

기존에도 RSS Aggregator에 들어오는 신규 정보를 개인의 선호에 맞게 분류하려는 연구가 있었다[6]. 사용자는 RSS Aggregator에 자신이 원하는 단어를 넣으면, 시스템은 RSS 문서내의 Item.title 과 Item.description에서 연관 단어를 찾아 분류해 내는 방식이다. 하지만 이는 아래와 같은 한계점을 지닌다.

첫째, 사용자가 입력한 단어와 연관된 단어를 찾기 위해, 별도의 Taxonomic Knowledge Base를 필요로 한다. 따라서 이미 잘 구축된 TKB에 있다는 전재로 하고 있는 것이다.

둘째, RSS 문서 내부에 존재하는 Item.title 이나 Item.description은 실제 문서의 제목과 해당 내용의 첫줄 정도에 밖에 지나지 않는다. 따라서 실제 문서가 사용자가 원하는 내용을 가지더라도 RSS 문서에 해당 단어가 위치하지 않는다면 분류해 낼 수 없다.

본 논문에서 제시하는 방식은, 분류를 하기 위해 RSS문서내의 Item.link 정보를 이용하여 실제 문서에 접근하여 처리 한다. 따라서 RSS문서 내의 정보에 국한되지 않는다. 또한 이미 분류된 초기 문서들에서 분류를 위한 정보를 추출해 내므로 별도의 TKB를 필요로 하지 않는다.

2.2 온톨로지를 이용한 문서 분류

최근 들어 온톨로지를 이용하여 개념적으로 문서를 분류하려는 연구가 많이 진행되고 있다. 온톨로지에 대한 정의는 다양하게 존재한다. 그중 가장 일반적으로 사용되는 것은, "온톨로지는 공유되는 개념에 대한 명시적 기술"이라는 정의이다[7].

온톨로지는 다양한 목적으로 구축되고 있다[8]. 이중 분류의 목적으로 온톨로지를 사용하여, 문서를 분류하고 응용하는 연구가 있다[9,10,11].

온톨로지 기반 추천 에이전트 연구[9]에서는 온톨로지를 "특정 주제에 대한 간단한 규칙들이나 의미적 연관관계와 단어들을 포함한 지식 용어들의 집합"으로 정의하고, 사용자가 이용하는 웹페이지를 온톨로지에 매핑시킴으로써 의미적으로 웹페이지를 구분하려고 한다.

SEWeP[10,11] 또한 온톨로지의 특징정보를 이용하여 문서를 분류하고, 이를 통하여 사용자들의 웹 내비게이션 패턴을 인식하고, 해당 사용자에게 차후 페이지를 추천한다.

이러한 연구들에서 사용된 온톨로지는 각 노드를 나타낼 수 있는 특징 단어 선별 방법이 필요하다. 따라서 본 논문에서 제시하는 방법을 이용하여 분류를 위한 온톨로지의 구성에 적용할 수 있다.

2.3 기타 문서 분류

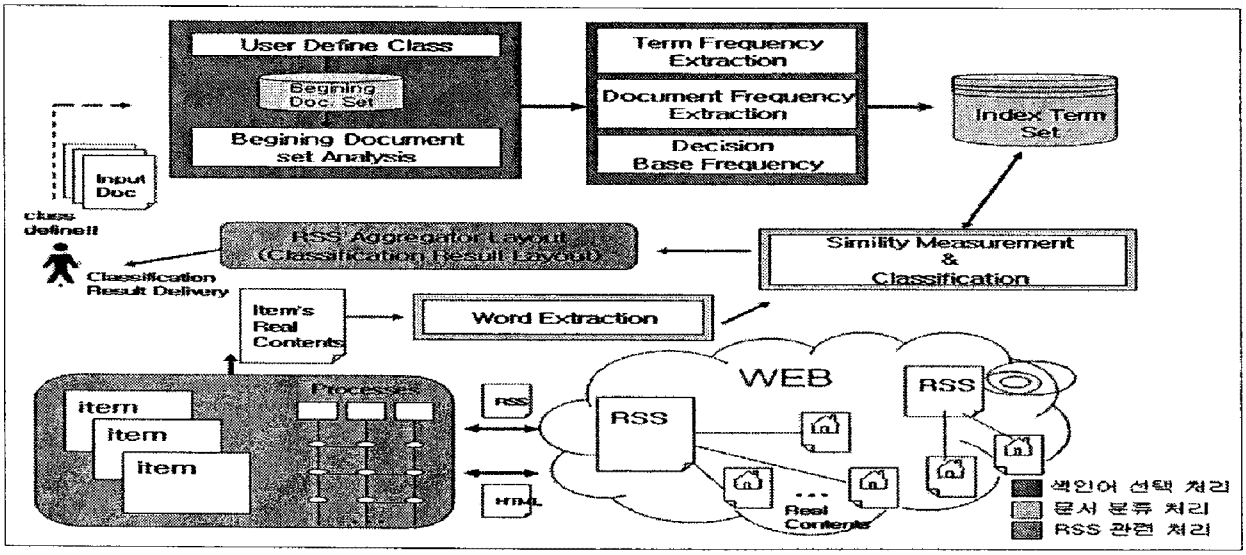
일반적으로 문서 분류에 좋은 성능을 나타내는 기계학습 방법으로 SVM(Support Vector Machine)이나 Naive Bayes Classifier 과 관한 연구들이 존재한다[12,13]. 하지만 이 방법들은 분류를 위한 색인어 추출 시 많은 시간을 요구한다. 따라서 새로운 분류를 위한 색인어를 구성하려면 또다시 학습의 시간이 필요하게 된다.

웹 정보는 고정적이지 않고, RSS의 특성상 새롭게 생성되는 정보들은 시대적 이슈에 해당하는 단어들을 포함하고 있다. 이는 계속 변하게 되므로 이를 적절히 분류하기 위해서는 분류를 위한 색인어들을 계속 갱신해 주어야 한다. 따라서 빠른 분류 색인어 집단 구성이 필요하다.

본 논문에서는 좀 더 간단한 방법으로, 빠른 시간 내에 분류를 위한 색인어 추출 집단을 구성할 수 있는 방법을 제시한다.

3.사용자에게 적합한 웹문서 전달 시스템

본 시스템은 웹에서 주기적으로 RSS를 통해 신규 자료를 얻어온다. 사용자는 이 시스템에 관심 분야를 생성하고, 해당 분야의 관련 문서들을 넣어준다. 시스템은 각 분야의 입력 문서들을 분석하여 분야별 색인어 집단을 추출한다. 시스템은 이 색인어 집단을 이용하여 RSS의 통하여 들어오는 신규 자료를 각각의 분야로 분류하여 사용자에게 전달한다.



<그림 2> 시스템 구조도

3.1 시스템 구성

첫째로, RSS 처리 모듈은 원격 사이트에서 RSS 문서를 가져와서 이전에 가져온 RSS 문서와의 비교를 통해 신규 자료 여부를 판단해 낸다.

둘째로, 분류 모듈이다. 여기서는 RSS 문서 내의 각 Item의 실제 문서에 접근하여 정보를 수집한다. 추출된 정보는 이전에 분류된 색인어 집단과 유사도 비교를 통하여 문서의 분야별 분류를 수행한다.

마지막으로, 분야별 색인어 선택 모듈이다. 이곳에서는 사용자가 입력한 이미 분류된 문서들을 분석하여, 각 분야별 색인어 집단을 선정한다. 이 과정을 통하여 고정 색인어 DB없이도 문서를 분류할 수 있게 된다. 이 시스템의 구조도는 <그림 2>와 같다.

3.2 RSS 문서 처리

RSS 문서의 처리를 위해 현재 별도의 RSS Aggregator를 구현하였으며, 동작 순서는 <그림3>과 같다. RSS Aggregator가 시작되면, 등록된 채널 정보를 이용하여 URL객체를 생성한다. 생성된 URL 객체를 이용하여 RSS 파일을 가져와 Xerces Parser를 이용하여 파싱한 후, 채널 객체로 만들어 필요한 앨리먼트를 추출한다. 만약 기존 채널 객체가 파일로 존재할 경우, 채널객체를 복원해 내어 현재 채널과 비교하는 과정을 통해 사이트의 신규 자료 여부가 판단된다.

3.3 웹 문서의 분류

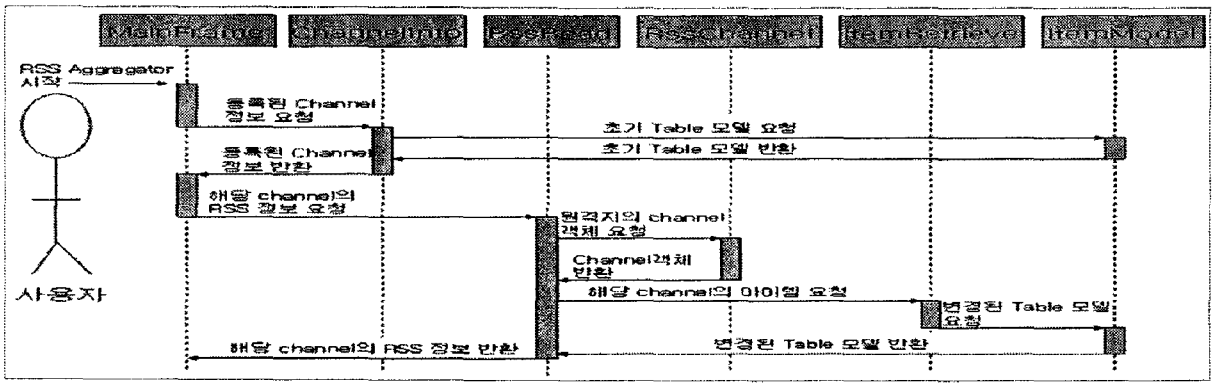
RSS 문서의 경우, 실제 문서의 링크 주소를 포함하고 있다. 따라서 해당 문서를 분류하기 위해 링크 주소에 존재하는 실제 문서를 가져와서 문서의 색인어를 추출하고, 시스템 내에 가지고 있는 색인어 집단과의 유사도 비교를 통해 분류를 하게 된다.

1) 분류를 위한 분야별 색인어 집단 선정

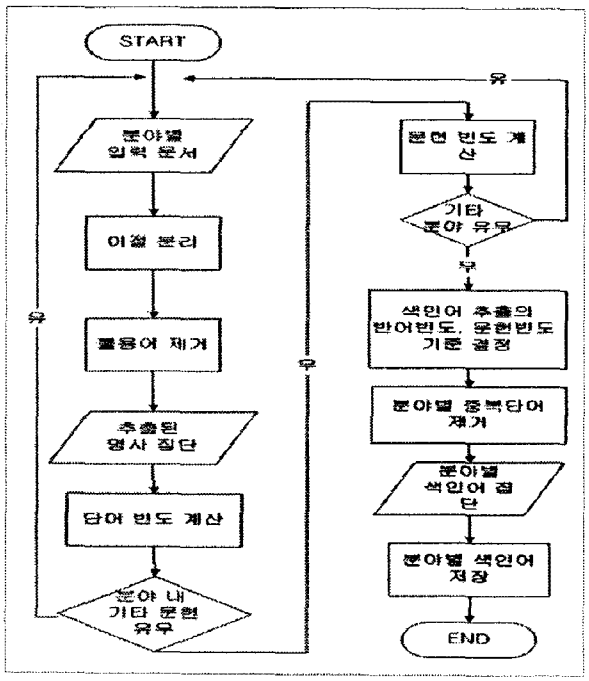
문서를 분류하기 위해선, 먼저 분류를 위한 색인어 집단이 필요하다. 이를 위해 사용자는 각 분야에 해당하는 문서들을 시스템에 초기 입력문서로 넣어준다. 시스템은 이 문서로부터 각 분야에 맞는 색인어들을 추출하여 분야별 색인어 집단을 생성하게 된다. 이 과정은 <그림4>와 같다.

2) 입력 문서의 색인어 추출

RSS 문서로부터 얻어낸 링크주소를 이용하여 가져온 실제 문서는 분류모듈에서 어절 단위 분리, 조사, 불용어 제거 작업을 통해, 문서 내 존재하는 모든 단어들을 추출하고 해당 단어의 빈도수를 기억한다. 추출된 모든 단어와 단어빈도수는 웹 콘텐츠 분류 시에 사용된다.



<그림 3> RSS Aggregator 동작 시퀀스 다이어그램



<그림 4> 분야별 색인어 추출 흐름도

3) 웹 문서 분류

입력 문서 내에서 추출된 단어들은 분야별 색인어 집단과의 유사도 계산을 통하여, 유사도가 가장 높은 분야에 분류된다.

유사도 계산식은 다음과 같다

$$Sim(C,d)=\frac{\sum_{i=1}^{N_c}(freq_{i,d}/\max_{l,d})}{N_c} \quad \cdots\cdots (1)$$

Nc: c분야 색인어 집단에서의 총 색인어 수

$freq_{i,d}$: 입력문서d에서 매칭 되는 단어i의 빈도수

$max_{l,d}$: 문서d에서 가장 자주 등장하는 단어의 빈도수

4. 실험

본 논문에서 실험을 위해 (정치, 경제, 문화, 국제, 사회, 스포츠, IT) 7개의 분야를 선정하고, 각 분야별로 이미 분류된 100개씩의 뉴스 문서를 준비하였다.

색인어 추출 시 고려해야 될 사항은 크게 세 가지이다. 첫째는 단어 빈도수(tf)이다. 한 문서 내에 몇 번 이상 나온 단어를 중요한 색인어로 취급할 것인가이다. 둘째는 문헌 빈도수(df)이다. 동일 분야에서, 해당 단어의 문헌 빈도수가 몇 번 이상이 그 분야를 대표할 수 있는 단어인지 결정하는 문제이다. 마지막은 분야별 중복 색인어 제거 기준이다. 분야를 대표하는 단어가 다른 분야에도 존재한다면 분류 시에 오류를 범하므로 이를 제거하는 기준을 정해야 한다. 만약 적절한 기준 없이 모두 제거해 버리면, 한 분야에서는 높은 문헌 빈도수를 가지는 단어가, 상대적으로 낮은 문헌 빈도수를 가지는 다른 분야의 단어에 의해 삭제될 가능성이 있다.

실시간 콘텐츠 분류를 위한 색인어 추출 시 궁극적 목적은, 최소한의 색인어 집단으로 최대한의 분류효과를 얻는 것이다. 색인어 집단의 크기가 작을수록, 유사도 비교 연산과정이 줄어듦으로 이는 실시간 분류 성능에도 큰 영향을 미친다. 따라서 분야를 대표하는 색인어 추출은 위 세 가지 요소의 최적 값을 선정하여 적절한 분류 성능을 가지는 최소한의 색인어 집합 선정을 해야 한다.

4.1 실험에 쓰인 자료 분석

1) 단어 빈도 분석

한 문서내의 모든 단어를 추출하여, 단어 빈도수를 분석한 결과는 다음과 같다.

75%~85% 정도의 단어가 빈도수 1을 가지며, 10%~20%정도가 빈도수 2를 가진다. 빈도수 3이상은 2%~5% 정도의 비중을 차지한다.

따라서 빈도수 1이상의 단어들로 분야의 색인어 집합을 구성하는 것과 빈도수 2이상의 단어들을 가지고 분야의 색인어 집합을 구성하는 것은 적게는 4배에서 많게는 10배 정도의 크기차이를 가진다.

2) 분야별 문헌 빈도 분석 및 색인어 크기

분야별로 단어빈도수 1이상/2이상/3이상 각각의 경우의 단어집합을 만들고, 분야별 문헌빈도수를 조사해 보았다. 예를 들어 <표2>의 (단어빈도1이상, 문헌빈도1이상)에서 "스포츠" 분야를 보면, 평균 문헌 빈도가 1.802로 나와 있다. 이것의 의미는 "스포츠"분야의 100개의 문서에서 빈도수 1이상의 모든 단어를 추출하고, 이 단어들의 문헌 빈도수를 구하였을 때, 평균 문헌 빈도수가 1.802라는 것이다.

<표2>에서 (단어빈도2이상, 문헌빈도1이상)의 "사회"분야를 보면 평균문헌 빈도수는 1.265이고, 색인어의 크기는 단어 1266개로 구

<표 2> 단어 빈도수 별 문헌 빈도 1이상의 최대/최소/평균값 과 분야별 색인어 수

			정치	경제	문화	국제	사회	스포츠	IT
단어 빈도1 이상	문헌 빈도1 이상	최대빈도	57	41	55	38	57	66	66
		최소빈도	1	1	1	1	1	1	1
		평균빈도	1.763	1.595	1.459	1.616	1.449	1.802	1.649
	색인어 수		7505	6712	10018	6481	7631	7553	7047
단어 빈도2 이상	문헌 빈도1 이상	최대빈도	29	14	12	24	10	34	20
		최소빈도	1	1	1	1	1	1	1
		평균빈도	1.720	1.454	1.329	1.563	1.265	1.803	1.522
	색인어 수		1109	1153	1578	923	1233	934	1158
단어 빈도3 이상	문헌 빈도1 이상	최대빈도	23	7	8	17	8	23	12
		최소빈도	1	1	1	1	1	1	1
		평균빈도	1.660	1.351	1.235	1.537	1.149	1.751	1.430
	색인어 수		409	444	605	361	482	334	440

<표 3> 단어 빈도수 별 문헌 빈도 2이상의 최대/최소/평균값 과 분야별 색인어 수

			정치	경제	문화	국제	사회	스포츠	IT
단어 빈도1 이상	문헌 빈도2 이상	최대빈도	57	41	55	38	57	66	66
		최소빈도	2	2	2	2	2	2	2
		평균빈도	4.183	3.646	3.412	3.768	3.606	3.952	3.804
	색인어 수		1798	1510	1907	1443	1316	2052	1631
단어 빈도2 이상	문헌 빈도2 이상	최대빈도	29	14	12	24	10	34	20
		최소빈도	2	2	2	2	2	2	2
		평균빈도	3.681	2.959	2.847	3.574	2.739	4.0	3.108
	색인어 수		298	267	281	202	188	250	287
단어 빈도3 이상	문헌 빈도2 이상	최대빈도	23	7	8	17	8	23	12
		최소빈도	2	2	2	2	2	2	2
		평균빈도	3.547	2.773	2.560	3.771	2.532	3.885	2.783
	색인어 수		106	88	91	70	47	87	106

<표 4> 단어 빈도수 별 문헌 빈도 3이상의 최대/최소/평균값 과 분야별 색인어 수

			정치	경제	문화	국제	사회	스포츠	IT
단어 빈도1 이상	문헌 빈도3 이상	최대빈도	57	41	55	38	57	66	66
		최소빈도	3	3	3	3	3	3	3
		평균빈도	6.186	5.619	5.299	5.796	5.614	5.804	5.924
	색인어 수		938	687	816	672	585	1053	750
단어 빈도2 이상	문헌 빈도3 이상	최대빈도	29	14	12	24	10	34	20
		최소빈도	3	3	3	3	3	3	3
		평균빈도	5.408	4.246	4.587	5.180	4.182	5.676	4.972
	색인어 수		147	114	94	100	64	136	107
단어 빈도3 이상	문헌 빈도3 이상	최대빈도	23	7	8	17	8	23	12
		최소빈도	3	3	3	3	3	3	3
		평균빈도	5.280	4.194	3.700	5.543	4.273	6.000	4.184
	색인어 수		50	31	30	35	11	41	38

성되었다는 것이다.

<표3>은 추출된 단어 중 색인어 집합을 선정할 때, 문헌빈도수 1인 것은 제외하고 생각하겠다는 것이다. 한 분야를 대표하는 단어가 되려면 최소한 2번 이상의 문헌 빈도수를 가져야 한다는 가정에서 조사된 것이다.

예를 들어, <표3>에서 (단어빈도2이상, 문헌빈도2이상)의 "국제"분야를 보면, "국제"분야의 100개의 문서에서 단어빈도수 2이상의 모든 단어를 취하고, 이 단어들을 대상으로 문헌빈도수를 조사하였을 때, 문헌빈도수 1이 나오는 것은 제외하고 색인어 단어로써 선정하는 것이다. 이 단어를 대상으로 평균 문헌 빈도수를 구해보면 3.574가 나온다.

마찬가지로 <표4>는 문헌빈도수가 3번 이상 나오는 것을 대상으로 조사한 수치이다.

3) 분야별 중복 제거 기준

단어 빈도수와 문헌 빈도수의 결정하여 분야별 색인어 집단 정하고 나서 고려해야 할 것은, 다른 분야와의 색인어 중복이다. 따라서 타 분야의 색인어로 선정된 것은 해당 분야에서 제거시켜야 분야별로 분류 시 효율성을 가지는 색인어 집단이 된다.

하지만 모든 중복 색인어를 제거할 시에는 문제점이 발생할 수 있다. 예를 들어, "정치" 분야에서 '대통령'이라는 단어의 문헌 빈도수가 38이고, "문화" 분야에서 문헌 빈도수가 2라고 한다면, 문헌빈도수 2이상의 것으로 색인어 집단을 구성하였을 시에 두 분야 모두 '대통령'이라는 중복 색인어를 가진다. 만약 이 두 영역에서 모두 '대통령'이라는 색인어를 제거한다면 "정치" 영역을 잘 설명할 수 있는 색인어를 버리는 것이 되므로 성능 저하의 요인이 될 수 있다.

이 경우, 한 가지 해결책으로 색인어 집단 선정 문헌 빈도수를 높이는 경우를 생각해 보자. 색인어 집단을 문헌 빈도수를 3이상인 것으로 구성한다면 앞의 문제는 해결된다. 하지만 또 다른 문제가 발생한다. 만약, '경기'라는 단어가 문헌 빈도수 6으로 "경제"영역에 색인어로 선정되었고, 같은 단어가 빈도수 7로 "스포츠"영역에 선정된 경우를 보면, 이 두 단어는 두 영역에서 모두 제거되지 않으므로 문서 분류 시 두 분야 간 문서 분류 시 성능 저하 요인이 된다. 이 해결책으로 각 분야 당 중요 문헌 빈도수를 고려하는 방법을 시도하였다.

<표2><표3><표4>에서 보여 지는 것과 같이 색인어 집단을 선정하고, 그 집단 내의 문헌 빈도 평균값을 이용한다. 각 분야 당 중요 문헌 빈도수는 분야 당 색인어 집단 내의 문헌 빈도 평균값의 정수 근삿값을 취한 것으로 이용한다. 따라서 분야별 중복 제거 기준은, 해당 단어의 문헌빈도수 차가 해당 분야의 중요 문헌 빈도수보다 클 때는 남기고, 그 이하일 때 제거하는 방법을 취하였다.

예를 들어, 분야A의 색인어k가 분야B의 색인어k에 의해 제거되지 않을 경우는 다음과 같다.

$$df_{k,A} - df_{k,B} > AVG(df_A) \quad \cdots\cdots \quad (2)$$

$df_{k,A}$: A분야에 속한 색인어 k
$df_{k,B}$: B분야에 속한 색인어 k
$AVG(df_A)$: A분야의 문헌 빈도수의 평균값

만약 식(2)을 적용하여 A분야에서 단어 k가 제거되지 않았을 경우, B분야에서는 분야 간 평균문헌빈도수의 차를 취한 결과가 음의 값을 가지므로 B분야에서 단어 k는 제거된다.

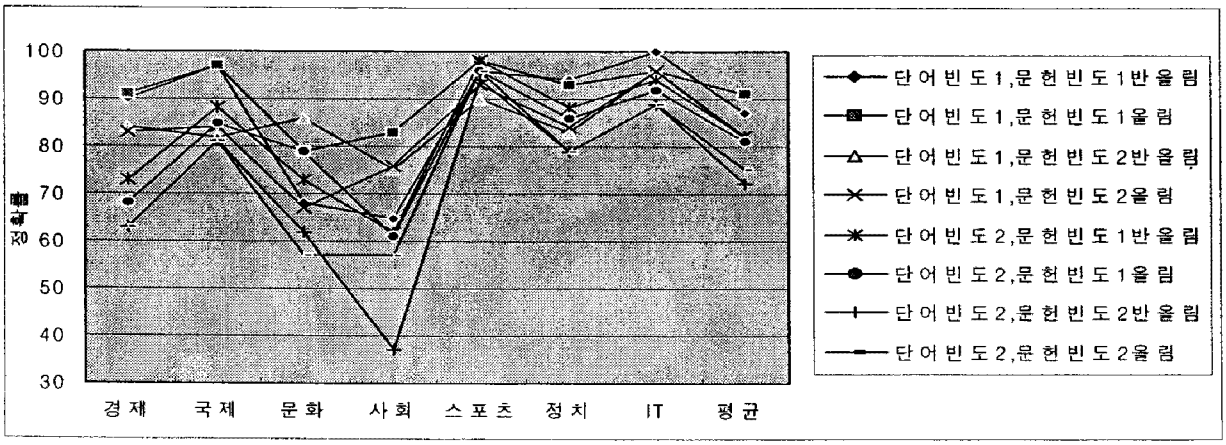
4.2 실험 결과

실험 결과, 단어 빈도수 3이상의 것만 색인어로 취하였을 때는 색인어 집단의 크기가 너무 작아져, 분류 성능이 현저히 저하되었다. 따라서 단어 빈도수 1이상과 2이상인 것을 대상으로 실험한 결과를 제시한다.

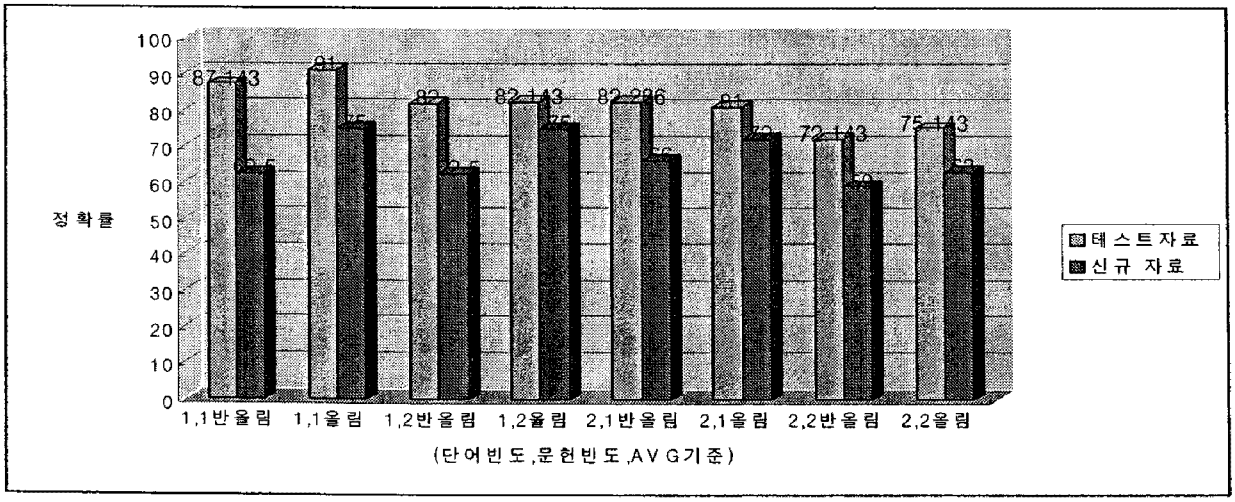
실험 시 분야별 중복 제거 기준에 사용된 평균 문헌빈도수 수의 최적 값을 찾기 위해 정수 근삿값을 반올림, 올림의 경우로 나누어 실험 한다. 먼저, 색인어 집단 추출 시 사용되었던 분야 당 100개의 문서를 넣고, 분류시켜 보았다.

<표5>에서 볼 수 있듯이, 색인어 집단 추출 시 사용되었던 문서이므로 꽤 높은 수치의 정확률을 가진다. '단어빈도수 1이상과 문헌빈도수 1이상의 평균값을 중요 문헌 빈도수'로 가질 때의 정확률이 90% 정도로 나왔고, '단

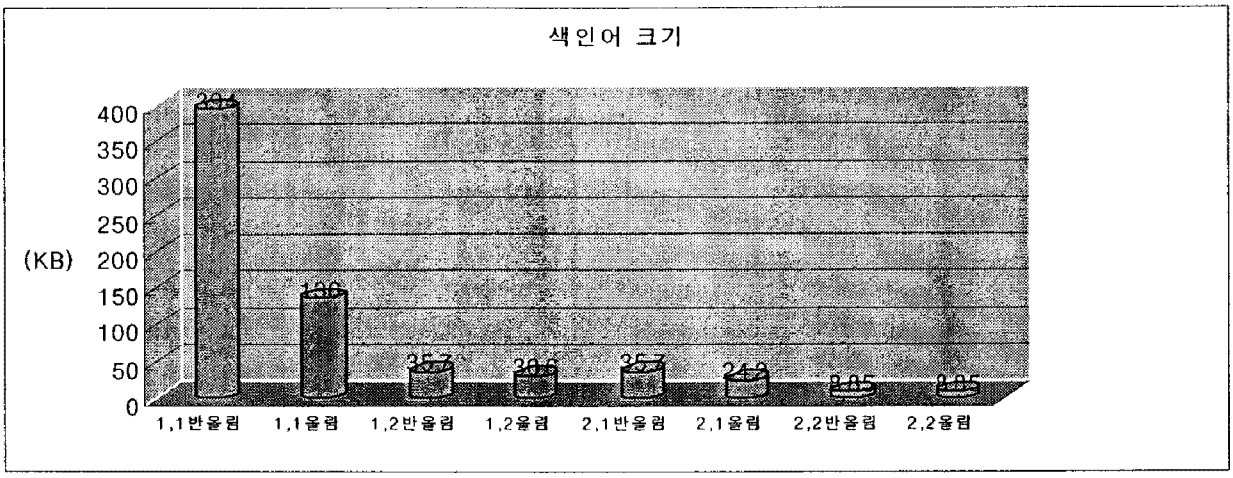
<표 5> 색인어 집단 추출에 사용된 분야별 100문서의 분류 정확률 측정 결과



<표 6> 색인어 집단 추출 시험 테스트 자료와 신규 자료의 정확률 비교



<표 7> 색인어 집단의 크기 비교



어빈도수 1이상과 문헌 빈도수 2이상의 평균값을 중요 문헌 빈도수'로 가지는 경우와 '단어 빈도수 2이상과 문헌 빈도수 1이상의 평균값을 중요 문헌 빈도수'로 가지는 경우는 80%를 약간 넘는 정확률을 보였다. 그리고 대체적으로 평균값에서 반올림보다는 올림을 취했을 때 정확률이 좀 더 우수했다.

이후에는 색인어 집단 추출 시 사용되었던 분야 당 100개의 자료 이외에, 임의의 문서 32개의 분류 실험을 하였다. <표6>에서 보듯이 기존 실험에 비해서는 정확률이 조금 떨어진다. 이때에는 '단어빈도수 1이상과 문헌 빈도수 1이상의 평균값(올림)을 중요 문헌 빈도수'로 취한 경우와 '단어빈도수 1이상과 문헌 빈도수 2이상의 평균값(올림)을 중요 문헌 빈도수'로 취한 경우, 그리고 '단어빈도수 2이상과 문헌 빈도수 1이상의 평균값(올림)을 중요 문헌 빈도수'로 취한 세 가지 경우의 약 정확률 75%로 선정되었다.

하지만, <표7>에서 알 수 있듯이, 위 세 경우의 색인어 크기는 394KB, 30.6KB, 24.2KB로 차이가 난다. 색인어 집단의 크기는 문서 분류 시 시간에 영향을 미치는 요인이다. 따라서 비슷한 성능을 가지는 세 가지 색인어 집단 중에서 '단어빈도수 2이상과 문헌 빈도수 1이상의 평균값(올림)을 중요 문헌 빈도수'로 하는 색인어 집단을 선정하는 것이 가장 빠른 문서 분류를 할 것이다. 결과를 정리해보면, 적정 정확률을 유지하며, 빠른 분류를 위한 최소 색인어 집단을 선정하는 방법은 <표8>과

<표 8> 색인어 집단 추출 시 빈도수/제약조건 구성 방법

(단어빈도수, 문헌빈도수)
$\lceil AVG(tf_c) \rceil$, $\lceil AVG(df_c) \rceil$
(중복 색인어 제거 기준)
Not Elimination, *where*
$df_{k,c1} - df_{k,c2} > df_c$

같이, 각 분야의 평균 단어빈도수와 평균 문헌 빈도수 평균의 올림 값을 취한 색인어 집단을 구성하고, 색인어 제거 기준을 적용한 방식을 취하는 것이 효율적이다.

5. 결론

RSS와 같은 신디케이션 기술은 사용자에게 신규 정보 여부를 알려줌으로써, 사용자의 불필요한 웹서핑 수고를 덜어준다. 하지만, 제대로 분류되지 못한 많은 양의 신규 정보는 또다시 사용자로 하여금 자신이 원하는 정보를 찾기 위한 검색을 요구한다.

따라서 본 논문에서는, 신규 자료를 얻어올 수 있는 RSS Aggregator 기능과 색인어 DB 없이도 적용할 수 있는 웹 문서 분류 방법을 적용하여, 사용자에게 효율적인 정보 전달을 할 수 있는 시스템을 제시하였다. 또한 분류를 위한 색인어 구성 시 이용할 수 있는 기준을 결정하기 위해서 실험을 통해 적정한 수치를 알아보았다. 사용자는 스스로 분야를 정하고,

관련 문서를 시스템에 넣어주면, 해당 분야의 색인어 집단이 추출되고, RSS를 통하여 들어오는 정보는 이를 기준으로 분류되어 사용자에게 전달된다. 이를 통해 사용자는 자신이 원하는 정보를 좀 더 효율적으로 전달 받을 수 있다. 본 논문의 테스트 자료는 뉴스의 7개영역으로 실험하였지만, 이는 다른 분야의 자료를 가지고 적용이 가능하다. 사용자는 자신의 관심분야의 문서들을 색인어 추출 시 사용되는 입력 문서로 사용하고, 이를 통해 개인화된 학습서비스 체계를 구축할 수 있다.

향후 과제로, 시대적 이슈가 되는 단어나, 새롭게 만들어져 사용되는 용어를 담고 있는 문서를 제대로 분류하기 위해, 초기 색인어 집단을 주기적으로 갱신해 줄 필요가 있으므로, 적절한 갱신 주기를 결정하는 연구가 필요하고, 중복 영역의 색인어들의 활용 방안을 모색해야 할 것이다.

6. 참고문헌

[1] PEW INTERNET & AMERICAN LIFE PROJECT, www.pewinternet.org, 2004.

[2] An Introduction to RSS for Educational Designers, www.downers.ca/files/RSS_Educ.htm, 2002.

[3] Weblogg-ed, www.weblogged-ed.com, 2005.

[4] World Wide Web Consortium, www.w3c.org, 2005.

[5] RSS Technology Reports, www.oasis-open.org/cover/rss.html, 2005.

[6] 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.

[7] Abasolo, J.M., "MELISA. An Ontology-based agent for Information retrieval in medicine." ECDL 2000 Workshop on the Semantic Web, pp.73-82. 2000.

[8] 최호섭, "정보검색 시스템과 온톨로지", 정보과학회논문지: 정보검색 제22권 제4호. pp.62-71, 2004.

[9] 정현섭, "개인화 된 웹 네비게이션을 위한 온톨로지 기반 추천 에이전트", 정보과학회논문지: 소프트웨어 및 응용 제30권 제1호, pp.40-50, 2003.

[10] Magdalini Eirinaki, "Web Personalization Integrating Content Semantics and Navigational Patterns", Proc. of the ACM WIDM'04, pp.72-79, 2004.

[11] M. Eirinaki, "SEWeP: Using Site Semantics and a Taxonomy to Enhance the Web Personalization Process",Proc. of the ACM SIGKDD '03, pp.99-108, 2003.

[12] McCallum, A. "A Comparison of Event-Models for Naive Bayes Text Classification", AAAI/ICML Workshop on Learning for Text Classification, pp.41-48, 1998.

[13] T. Mitchell. 1997. Machine learning. The Mc Graw Hill Companies. Inc. 2005.