TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법

Keyword Extraction from News Corpus using Modified TF-IDF

이성직(Sungjick Lee)*, 김한준(Han-joon Kim)**

초 록

키워드 추출은 정보검색, 문서 분류, 요약, 주제탐지 등의 텍스트 마이닝 분야에서 기반이 되는 기술이다. 대용량 전자문서로부터 추출된 키워드들은 텍스트 마이닝을 위한 중요 속성으로 활용되어 문서 브라우징, 주제탐지, 자동분류, 정보검색 시스템 등의 성능을 높이는데 기여한다. 본 논문에서는 인터넷 포털 사이트에 게재되는 대용량 뉴스문서집합을 대상으로 키워드 추출을 수행하여 분야별 주제를 제시할 수 있는 키워드를 추출하는 새로운 기법을 제안한다. 기본적으로 키워드 추출을 위해 기존 TF-IDF 모델을 고찰, 이것의 6가지 변형식을 고안하여 이를 기반으로 각 분야별 후보 키워드를 추출한다. 또한 분야별로 추출된 단어들의 분야간 교차비교분석을 통해 불용어 수준의 의미 없는 단어를 제거함으로써 그 성능을 높인다. 제안 기법의 효용성을 입증하기 위해 한글 뉴스 기사 문서에서 추출한 키워드의 질을 비교하였으며, 또한 주제 변화를 탐지하기 위해 시간에 따른 키워드 집합의 변화를 보인다.

ABSTRACT

Keyword extraction is an important and essential technique for text mining applications such as information retrieval, text categorization, summarization and topic detection. A set of keywords extracted from a large-scale electronic document data are used for significant features for text mining algorithms and they contribute to improve the performance of document browsing, topic detection, and automated text classification. This paper presents a keyword extraction technique that can be used to detect topics for each news domain from a large document collection of internet news portal sites. Basically, we have used six variants of traditional TF-IDF weighting model. On top of the TF-IDF model, we propose a word filtering technique called 'cross-domain comparison filtering'. To prove effectiveness of our method, we have analyzed usefulness of keywords extracted from Korean news articles and have presented changes of the keywords over time of each news domain.

키워드 : 텍스트 마이닝, 정보검색, 키워드추출, 주제탐지, TF-IDF
Text Mining, Information Retrieval, Keyword Extraction, Topic Detection, TF-IDF

본 연구는 2008년 서울형산업 기술개발 지원사업(NT080624)의 지원을 받았으며, 또한 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성·지원(NIPA-2009-C1090-0902-0031)의 연구결과로 수행되었음.

* 서울시립대학교 전자전기컴퓨터공학부
** 교신저자, 서울시립대학교 전자전기컴퓨터공학부

2009년 09월 22일 접수, 2009년 10월 06일 심사완료 후 2009년 11월 06일 게재확정.

1. 서 론

정보통신부와 한국인터넷진흥원이 실시한 '2007년 하반기 정보화 실태 조사'에 따르면, 인터넷 이용자의 과반수(58.5%)가 포털 사이트의 뉴스서비스를 이용하고 있다[2]. 이용자가 많은 포털 사이트 중 하나인 네이버(www.naver.com)에는 하루 10,000개 이상의 뉴스 기사가 게재되고 있지만, 이러한 대용량의 뉴스 기사를 짧은 시간 내에 파악하는 것은 쉽지 않다. 그래서 포털 사이트들은 이슈가 되고 있는 뉴스를 선택적으로 먼저 보여줄 필요가 있으며, 이를 위해 키워드 추출 기술(Keyword Extraction)이 유용하게 사용될 수 있다.

키워드 추출은 정보검색(Information Retrieval), 문서분류(Text Categorization), 주제탐색(Topic Detection), 문서요약(Document Summarization) 등을 포함한 텍스트 마이닝(Text Mining) 분야의 연구에서 주요 속성(Feature) 추출을 위해 사용되는 기술이다. 일반적으로, 문서 내부에 존재하는 단어의 중요도를 평가하기 위해서 현재 많은 검색엔진이 채택하고 있는 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 모델을 사용할 수 있다[6, 7, 8]. TF-IDF모델은 벡터 공간모델(Vector Space Model) 기반 정보검색을 위해서 문서를 표현하는 원리이며, 기본적으로 개별 문서에서 각 단어의 상대적 중요도를 표현할 수 있어서 개별 문서에 존재하는 키워드를 추출하는데 활용하고 있다. 본 연구의 키워드 추출은 개별 문서 범위가 아닌 '문서 집합 전체' 범위를 가정하기 때문에 기존 TF-IDF 모델의 원리를 유지하면서 추출 범위를 고려한 변형식을 제안한다. 즉 본 논문에서는 주어진 뉴스문서 집합 전체 범위에서 키워드를 추출하기 위해 6가지의 수정된 TF-IDF 가중치 모델과 이를 통해 얻은 키워드 집합을 한층 더 개선하기 위해 분야별 후보 키워드 집합을 통계적으로 교차비교하는 기법을 소개한다. 본 제안 기법에 의해 추출된 키워드는 뉴스 문서 집합의 요약 정보를 보여 줄 수 있으며, 이에 따라 뉴스 기사의 분류, 효율적인 뉴스 탐색 등에 활용될 수 있을 것으로 평가한다.

2. 관련 연구

인터넷상의 전자문서의 증가와 함께 주제탐사를 목적으로 한 키워드 추출 연구가 활발히 수행되고 있다. 키워드 추출을 위해 [4]에서는 단어들이 동시에 출현하는 통계적 정보를 활용하였으며, [9]에서는 인터넷 검색을 위한 색인 생성에 사용되는 PageRank 알고리즘을 이용하였다. 또한 기계학습(Machine Learning)의 원리를 적용한 연구가 있으며, [10]에서는 Support Vector Machine 알고리즘을, [3]에서는 Neural Network Model 알고리즘을 이용하였다. 기계학습 알고리즘을 이용한 방법은 키워드 추출을 위한 예측모델을 만들기 위해 고품질의 학습데이터를 준비해야 하는데 이에 대한 비용이 매우 클 뿐만 아니라, 그것의 정확도가 학습데이터에 의존하므로 수시로 변화하는 뉴스 문서에 이를 적용하는 것은 바람직하지 못하다. 기본적으로 키워드 추출은 실용적인 관점에서 비감독형 통계기법이 바람직하며, TF-IDF 가중치 모델은 이에 준하는 효과적인 방법이라 할 수 있다. 또

한 기존의 연구는 키워드 추출의 대상을 주로 단일 문서로 한정하는 반면에, 본 논문에서는 뉴스 문서 전체 집합을 대상으로 하여 비교적 비용이 적게 드는 방법으로 키워드를 추출했다는데 그 의의가 크다.

3. 키워드 추출 방법

본 연구의 '키워드 추출' 문제는 뉴스 문서 집합에서 각 분야(예를 들면, '정치', '경제', '사회' 등)의 주요 키워드를 찾는 것으로 정의한다. 각 분야별로 현재 화제가 되고 있는 키워드를 추출함으로써 뉴스 요약 및 뉴스 경향을 파악할 수 있는 기초를 만들 수 있다. 특정 분야의 뉴스 문서집합에서 키워드를 골라 내기 위한 기본적인 아이디어는 각 단어의 가중치를 계산한 후, 상위 적당 개수의 가중치를 가지는 키워드를 선정하는 것이다.

본 연구에서 제안하는 키워드 추출 기법은 2단계로 이루어진다. 첫 단계로서, 전체 문서 집합에 존재하는 단어를 정의된 가중치로 정렬하여 그 값이 주어진 임계값 보다 큰 '후보 키워드'들을 골라낸다. 그 후보 키워드 집합은 적정 수준의 키워드를 포함하고는 있지만, TF-IDF의 한계로 인해 불용어(Stopword) 수준의 단어(예를 들면, '기자', '신문사')가 포함될 수 있다. 그래서 두 번째 단계로서, 각 분야에서 얻어진 후보 단어들의 순위를 교차비교함으로써 각 분야의 대표단어로서의 키워드 집합을 얻게 된다. 본 절에서는 키워드 추출이 근간이 되는 TF-IDF 가중치 모델과 그것의 6가지 변형, 그리고 키워드의 분야간 교차비교 기법을 소개한다.

3.1 키워드 추출을 위한 TF-IDF 가중치 모델의 변형

앞서 언급한 바와 같이, 본 연구의 목적은 개별 문서가 아닌 적정 문서집합으로부터 주요 키워드를 추출하는 것이다. 그래서 주어진 문서집합에서 출현 단어의 중요도를 측정하기 위해 TF-IDF 가중치를 변형하고자 한다. 우선 본래의 TF-IDF 가중치 모델을 살펴보기로 한다.

3.1.1 TF-IDF 가중치 모델

TF-IDF 가중치 모델은 정보검색 및 텍스트마이닝을 위해서 문서 내부의 단어간 상대적 중요도를 평가하기 위해 문서의 표현방식으로서 고안된 것이다. 이 TF-IDF가중치로 표현된 문서는, 정보검색엔진에서 주어진 질의어와 가장 유사한 문서들의 순위를 결정할 수 있게 할 뿐만 아니라, 유사 문서들의 그룹(또는 클러스터)을 찾는 문서군집화를 용이하게 한다. TF-IDF 값이 큰 단어는 그것이 속한 문서의 주제 또는 의미를 결정지을 가능성이 크며, 따라서 이 측정치를 주요 키워드를 추출할 수 있는 척도로 활용할 수 있다.

<표 1>에서 보는 바와 같이, TF-IDF 가중치는 TF(Term Frequency)값과 IDF(Inverse Document Frequency)값을 곱한 것이다. TF값은 한 문서 내에서 특정 단어가 출현한 빈도수를 의미한다. 이 값을 가중치 모델에 포함시키는 것은, 주어진 단어가 문서 내에서 많이 출현할수록 상대적으로 더 중요하다는 가정을 반영한 것이다. 실제적으로 활용되는 TF값은 문서 내부의 단어 출현 빈도를 모든 단어의 총 출현 회수로 나누어 정규화한 형

〈표 1〉 본래의 TF-IDF 가중치 모델

TF값	$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ $n_{i,j}$: 단어 ti가 문서 dj에서 출현한 회수 $\sum_k n_{k,j}$: 문서 dj에서 모든 단어가 출현한 회수
IDF값	$idf_i = \log \frac{\lvert D \rvert}{\lvert \{d_j \mid t_j \in d_j\} \rvert}$ $\lvert D \rvert$: 문서집합에 포함되어 있는 문서의 수 $\lvert \{d_j \mid t_j \in d_j\} \rvert$: 단어 t_j가 등장하는 문서의 수
TF-IDF 가중치	$TFIDF_{i,j} = tf_{i,j} \times idf_i$

태를 띤다. 이는 문서의 크기에 따른 TF값의 편중 현상을 방지하기 위해서이다. 그런데 이 TF값만을 가지고 문서를 표현하기에는 부당한 측면이 있다. 이는 TF값이 지나치게 큰 단어가 오히려 불용어 수준의 단어에 해당할 수 있기 때문이다. 이를 해결하기 위해 IDF 인자를 도입한다.

IDF 값은 문서 집합에 포함되어 있는 문서 수를 특정 단어가 나타난 문서의 수(Document Frequency)로 나눈 것이다. 이는 상대적으로 많은 문서에 출현한 단어의 IDF 값은 작게 되고, 반대로 한쪽으로 편중하여 나타난 단어의 IDF 값이 커짐을 의미한다. 이는 IDF 값이 작은 단어는 보편적인 단어일 가능성이 크며, 반대로 IDF 값이 큰 단어는 문서 내에서 주요 의미를 가지는 단어로 분별하기 위한 의도를 반영한 것이다. [5]에서는 IDF 수치의 이론적 정당성을 보여 준다. TF-IDF 값은 TF인자와 IDF 인자를 곱한 것이며, 이 TF-IDF 가중치에 입각하여 '문서 내부'에서 많이 출현하고, 전체 문서집합에서 출현하는 문서들의 수가 적은 단어가 중요한 단어로서 평가될 수 있다. 본 논문에서는 하나의 문서가 아닌 문서집합 수준에서 키워드를 추출하고자 하기 때문에, TF-IDF모델의 단어간 상대적 중요도를 평가하는 범위를 '개별 문서'에서 '전체 문서집합'으로 확장해야 한다.

3.1.2 TF-IDF 가중치 모델의 변형

주어진 문서집합으로부터 키워드 추출을 위한 TF값은 문서 내에서의 특정 단어의 출현 빈도가 아닌 문서집합 내에서의 출현빈도로 정의한다. 그리고 키워드 추출의 정확도 향상을 위해 두 가지 방식으로 TF값을 정규화한다(<그림 1>참조). 이때 IDF 인자는 본래의 형태를 유지한다. 그래서 TF-IDF 값을 구성할 때 TF값에도 로그를 취하는 정규화 형태를 추가한다. 결국 6가지의 TF-IDF 변형식을 얻을 수 있게 된다.

3.1.2.1 TF식의 수정

<그림 1>에서 보는 바와 같이, 뉴스 문서집합으로부터 키워드 추출을 위하여 <그림 1>의 TF 변형 식 세 가지, 즉 BTF, NTF1, NTF2

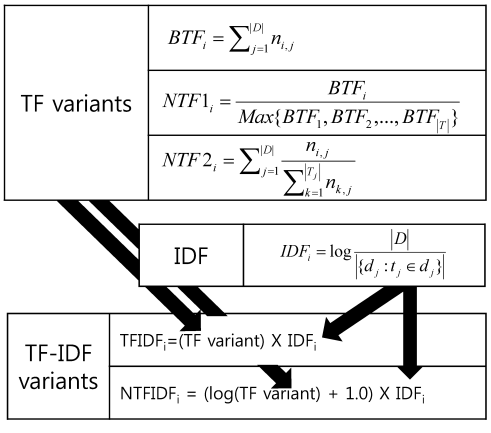
을 제안한다. 전체 문서집합 범위에서 키워드를 추출하는 것이 목적이기 때문에, 기본적으로 전체 뉴스문서에서 특정 단어가 출현한 회수를 더하여 그 값을 구하는 기본적인 방법을 사용하며, 이를 BTF(Basic Term Frequency)라 칭한다. 하지만 이는 각 문서에서의 출현빈도를 단순히 더한 값이어서 단어간 상대적 중요도를 나타내지 못할 가능성이 있다. 이를 보완하기 위하여, 두 가지 정규화된 TF식(Normalized Term Frequency)을 제안하며, 이를 각각 NTF1과 NTF2라 칭한다. NTF1은 BTF 값을 문서집합 내에서 출현한 모든 단어들의 BTF 값 중 최대값으로 나누어 정규화한다. 그리고 NTF2는 해당 단어의 문서에서의 발생빈도를 각 문서의 모든 단어에 대한 발생빈도로 나누어 더한 값으로 정의한다. 기본적으로 BTF 값이 큰 단어가 중요도가 높을 가능성이 크지만, 뉴스문서 길이가 일정하지 않은 경우에 공정하지 못하므로, NTF1과 NTF2와 같이 문서 길이가 달라서 생기는 가중치의 과도한 편차를 최소화하고자 한다.

3.1.2.2 TF-IDF의 조합

<그림 1>에서 보는 바와 같이, BTF, NTF1, NTF2 등의 TF 변형식을 사용하여 계산한 값과 IDF 식을 이용해 계산한 값을 사용해서 최종적으로 TF-IDF 변형식을 계산한다. 여기서 BTF와 NTF2의 식으로 계산한 TF값이 로그를 내재한 IDF값보다 상대적으로 매우 커질 수 있으므로, 두 인자가 유사한 범위의 값을 가지도록, TF값에도 로그를 취하여 IDF 값과 곱하는 방법이 필요할 수 있다. 그리고 BTF, NTF1, NTF2를 이용해 계산한 값이 1보다 작을 경우를 고려하여 로그를 취한 값에 '1'을 더하여 IDF 값과 곱한다. 이 계산식을 NTFIDF(Normalized TFIDF)라고 부르기로 한다.

3.2 분야간 단어 교차비교

각 분야별 뉴스 문서집합은 주어진 뉴스 문서집합의 한 부분집합이다. 따라서 전체 뉴스 문서집합 전체에서 높은 TF-IDF값을 가지는 단어들이 분야별 뉴스 문서집합에서도 높은 TF-IDF값을 가질 가능성이 있다. 문제는 그러한 단어들의 일부가 전체 뉴스기사를 대표한다기 보다는 불용어에 가까운 의미없는 단어일 가능성이 크다는 것이다. 예를 들어, 뉴스 문서의 작성자인 기자의 이름과 언론사의 이름은 거의 모든 뉴스 문서의 본문에 포함되어 있어 높은 가중치를 얻을 수 있다. 이런 단어들은 각 분야의 뉴스 문서집합에서 키워드로서 부적절하므로 제거되어야 한다.

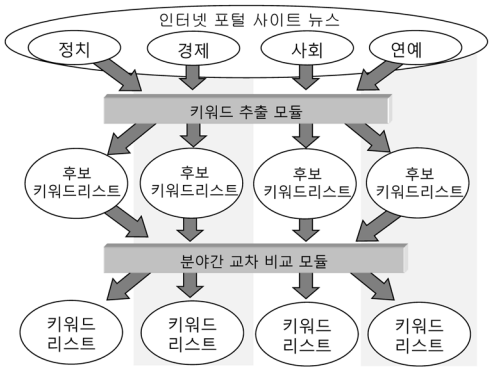


|D| : 특정분야 뉴스문서 집합에 포함된 문서의 수
|T| : 특정분야 뉴스 문서 집합에서 출현한 모든 단어의 수
$|T_j|$: 특정분야 뉴스에서 출현한 모든 단어의 수
$n_{i,j}$: 뉴스문서 j에서 단어 i가 출현한 회수

〈그림 1〉 키워드 추출을 위한 TF-IDF 모델의 수정

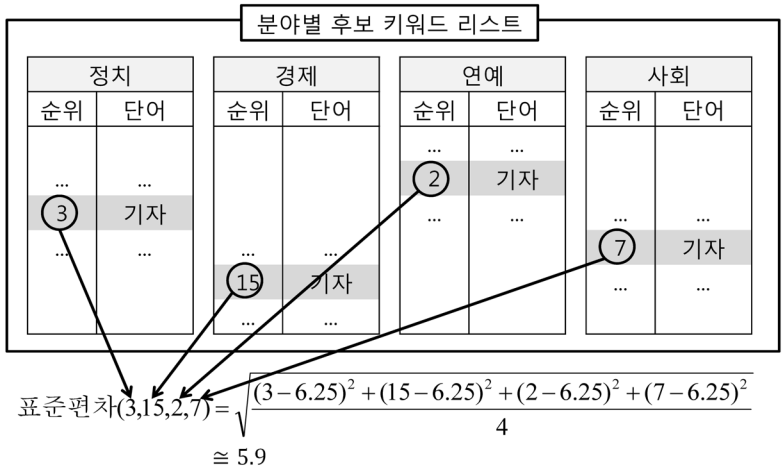
이를 위해 본 논문에서는 분야간 교차비교 기법을 제안한다. 본 논문의 키워드 추출 문제 문제는 자동문서분류(Text Classification) 분야에서의 클래스(Class)별 대표 속성(Feature)을 추출하는 문제와 유사하다. 특정 클래스에 소속된 문서의 주요 속성은 그 클래스의 개념을 명확하게 규정할 수 있는 단어(구)이어야 한다. 이와 유사하게, 각 뉴스분야에 출현하는 키워드는 그 분야의 특징을 잘 설명하고, 또한 타 분야와의 구분을 높이는 것이어야 한다.

그래서 본 연구에서는 각 분야에서 1차 생성한 후보 키워드들을 비교하여 동일 단어의 출현을 억제함으로써 키워드 선정의 정확도를 높이고자 한다. <그림 2>에서 보는 바와 같이, 첫 단계로서, *키워드 추출* 모듈이 인터넷 포털 사이트에 HTML 페이지 형식으로 게재되어 있는 뉴스 문서를 수집하여 분야별 후보 키워드집합을 생성한다. 두번째 단계로서, *분야간 교차비교* 모듈이 최종 키워드집합을 생성하게 된다.



〈그림 2〉 키워드 추출 시스템 및 과정

분야간 교차비교는 키워드 순위에 대한 표준편차 값을 계산함으로써 간단히 이루어질 수 있다. 만약 어떤 단어가 한 분야의 후보 키워드 집합에서 높은 TF-IDF 가중치를 가져 상위 순위에 있지만, 타 분야에서는 낮은 가중치를 받아 하위 순위에 있는 경우, 그 단어는 높은 가중치를 가지는 분야에서의 키워드로 선택된다. 반대로, 각 분야에서 TF-IDF 가중치가 비슷한 순위를 가지는 단어인 경우, 해당 분야의 성격을 규정하는 성질이 약하다



〈그림 3〉 분야간 교차비교의 예

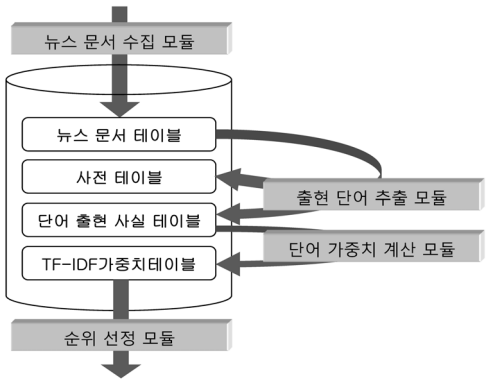
고 판단하여 삭제하게 된다.

다시 말해서, 각 분야의 뉴스문서집합으로부터 얻어진 후보 키워드집합에 대하여, 각 분야별 순위값을 계산한 후, 특정 단어들의 순위값들의 표준편차가 주어진 임계값 이하인 경우 키워드 선정에서 제외된다. <그림 3>은 이러한 계산 과정의 예를 보여준 것이다. 표준편차의 임계값을 10이라고 하자. 만약 '기자' 라는 단어가 각 분야의 후보 키워드 집합에 모두 등장하고, 각 분야에서 3위, 15위, 7위, 2위의 순위를 얻었다. 이 순위값들의 표준편차는 임계값 10을 초과하지 못하여, 뉴스 키워드로서의 역할을 못하는 것으로 판단하여 삭제하게 된다.

4. 시스템 구현

4.1 모듈 및 데이터베이스 설계

<그림 4>는 키워드 추출 모듈의 내부 구조를 보여준다. 키워드 추출을 위해 구축된 데이터 베이스는 뉴스문서, 사전, 단어출현사실, TF-IDF 가중치 등의 테이블을 포함한다. 수집된 문서는 뉴스문서 테이블에 저장하고, 이 테이블을 조회하여 출현 단어를 확인한 후에 사전 테이블과 단어출현사실 테이블에 관련 정보를 입력한다. 여기서 사전 테이블은 출현 단어를 중복 없이 관리하기 위한 것이고, 단어출현사실 테이블은 각 문서에서의 단어 출현 사실을 기록하기 위한 것이다. 이 단어출현사실 테이블을 기반으로 TF-IDF 가중치를 계산하여 그 결과를 TF-IDF 가중치 테이블에 계산 결과를 저장한다. 분야별 후보



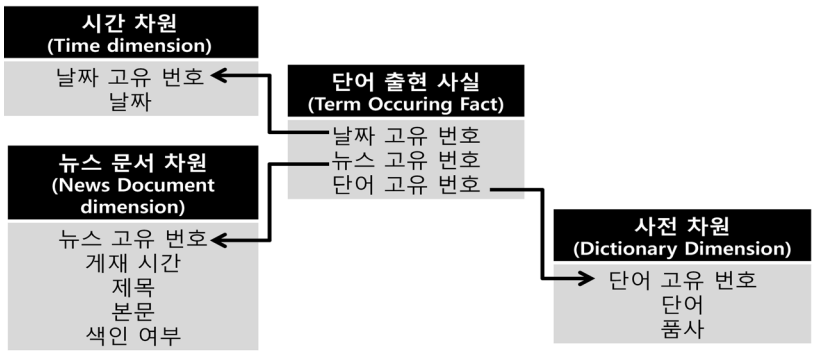
〈그림 4〉 키워드 추출 모듈의 내부 구조

키워드집합은 TF-IDF 가중치 테이블의 데이터를 분야별로 정렬함으로써 얻어지게 된다.

구축된 테이블 중에서 핵심이 되는 것은 단어출현사실 테이블로서, 이는 데이터 웨어하우스(Data Warehouse)의 구축에 활용되는 스타스키마(Star Schema)의 형태를 띤다(<그림 5> 참조). 중심에 위치한 사실(Fact) 테이블은 개별적인 사건의 발생 사실을 시간 순으로 기록하는 것이므로, 뉴스문서로부터 추출한 각 단어를 하나의 발생 사실로 간주하여 단어출현사실 테이블에 하나의 레코드로 삽입된다. 사실 테이블에 저장된 출현 단어 정보는 차원테이블로서 정의된 시간 테이블, 뉴스문서 테이블, 사전 테이블을 외래키를 통해 참조함으로써 특정 날짜에 어떤 단어가 어떤 문서에 출현하였는지를 알 수 있다.

4.2 단일 명사의 추출과 복합명사의 생성

한글의 경우, 단일명사는 그 의미가 불확실한 경우가 많다. 예를 들어, '여성 비례대표 의원들의 입법 활동이 두드러진 것으로 조사됐다.'라는 문장에서 '여성', '비례대표', '의원



〈그림 5〉 키워드 추출을 위한 스타 스키마(Star Schema)

〈표 2〉 복합 명사 생성에 사용되는 패턴

1. 조사가 생략된 명사/인접한 명사(예 : 형태소 분석)
2. 관형격 조사 결합 명사/피수식 명사(예 : 정보의 검색)
3. 목적격, 주격조사 결합 명사/서술형 명사(예 : 정보를 검색하는)
4. 관형화된 서술형 명사/피수식 내포문의 명사(예 : 정보를 처리하는 시스템)

들', '입법', '활동', '조사' 등과 같은 단일명사가 추출될 수 있다. 이와 같은 단어들 중에서 키워드를 선별한다면 그것이 속한 문장 또는 문서의 전반적 의미를 표현하기 어려울 것이다. 비교해서, 복합명사로서 '여성 비례대표 의원들', '입법 활동' 등이 추출된다면 이는 키워드로서의 의미성이 훨씬 높아질 것이다. 본 연구에서는 복합명사의 추출을 위해 <표 2>의 복합명사 생성 패턴을 이용한다. 본 실험에서는 KLT 라이브러리[2]를 이용하여 단일명사 추출과 형태소 분석을 수행한다[1]. 단일명사 추출은 KLT의 색인어 추출 함수를 이용하였으며, 이는 불용어 사전을 활용하여 의미 없는 단어를 제거해준다. 형태소 분석 기능은 단어의 품사 정보, 조사정보, 역할 정보 등을 반환한다. 이러한 정보를 바탕으로 <표 2>의 4가지 패턴에 매칭함으로써 복합명사를 결정하게 되며, 만약 단일명사로 추출된 단어가 복합명사 합성에 사용되었다면, 해당 단일명사의 출현 사실은 무시되고, 생성된 복합명사의 출현 사실만을 저장한다.

추출된 단일 명사와 복합 명사에 대해 우선 사전 테이블을 검색하여 동일한 명사가 없는 경우, 이를 사전 테이블에 넣고 고유번호를 부여한다. 그리고 출현한 문서와 날짜의 고유번호와 사전 테이블에 포함된 단어의 고유번호를 묶어 레코드를 생성하여 단어출현 사실 테이블에 입력한다.

5. 성능 평가

본 연구에서는 기존 TF-IDF 모델의 수정

2) KLT 라이브러리는 국민대학교 언어공학, 정보검색 연구실(http://nlp.kookmin.ac.k)에서 연구용으로 배포한 것이다.

과 뉴스 분야간 교차비교 분석을 이용하여 뉴스 문서집합에서 각 분야별 주요 키워드를 추출하는 기법을 제안하였다. 이 기법의 실제적 의미를 평가하기 위해 인터넷 포털사이트 네이버(http://www.naver.com)에 게재되는 뉴스 문서를 대상으로 하여 키워드 추출 실험을 수행하였다.[3] 네이버는 실시간 인터넷기사를 속보, 정치, 경제, 사회, 생활/문화, 세계, IT/과학, 연예, 칼럼, 영문, 매거진, 전문지 등으로 분류하여 제공하고 있다. 본 실험에서는 정치, 경제, 사회, 연예 등의 네 분야의 뉴스 기사를 실험 대상으로 하였다. 평일 하루 동안에 정치분야에서 약 1,500개, 경제분야에서 약 6,000개, 사회분야 약 4,000개, 연예분야 약 1,200개의 뉴스를 수집하여 각 분야별 뉴스 문서집합을 구성한다.

본 논문에서 제시한 뉴스 문서 집합에서의 키워드 추출 방법의 성능을 보이기 위해 다음과 같이 실험 결과를 제시한다. 3.1절에서는 전통적인 TF-IDF를 전체 문서 집합에 적용하기 위한 6가지 변형식에 따른 후보 키워드 리스트의 결과를 제시하고 비교한다. 3.2절에서는 분야간 교차비교의 성능을 보이기 위해 적용 전의 후보 키워드 리스트와 적용 후의 키워드 리스트를 보인다. 또한 분야간 교차비교를 할 때 적용되는 순위에 대한 표준편차값의 임계값에 따른 키워드 리스트의 변화를 보임으로써 임계값을 적절히 조절해야 함을 보인다. 마지막으로 시간에 따른 키워드의 변화를 보여 각 분야에서 화제가 되고 있는 주제가 시간에 따라 변화하는 양상을 보인다.

5.1 분야별 후보 키워드 리스트

<표 3>은 '정치' 분야의 뉴스문서집합을 대상으로 6가지의 TF-IDF변형식을 적용하여 얻은 후보 키워드 집합을 보여준다. 여기서, 기존 TF-IDF 가중치를 이용해 얻은 키워드는, 문서 내에서 각 단어에 기존 TF-IDF 가중치를 할당한 후, 전체 문서 집합에서 순위를 매긴 결과에서 상위 10개의 단어들이다. <표 3>에서 보는 바와 같이 기존 TF-IDF가중치를 통해 추출한 키워드들은 '한나라당'을 제외하고 대부분 불용어 수준에 가까운 것이다. 비교해서, NTF2식을 이용한 2가지 결과를 제외하고는 기존의 TF-IDF 조합보다 개선된 결과를 보여준다. 이는 일반적으로 게재되는 뉴스 기사의 길이가 짧아서 문서 내 TF값에 지나치게 영향을 받기 때문이다. 가장 효과가 좋은 것은 NTF1과 NTFIDF 변형식의 조합을 사용한 경우이다.

6가지 TF-IDF 변형식을 이용한 경우에도, '기자', '무단전재', '재배포 금지' 등의 의미 없는 단어가 포함되었으며, 더구나 이러한 단어들이 상위 순위에 위치하고 있다. 이는 이 단어가 문서집합 내의 대부분의 뉴스문서에서 출현하여 TF값과 TF-IDF 값이 매우 크기 때문이다. 비근한 예로서, 뉴스문서의 본문에는 작성자를 나타내는 'OOO 기자'와 같은 문자열이 포함되어 있어 대부분의 뉴스문서에서 '기자'라는 단어가 출현한다. 이러한 단어는 매우 큰 TF값을 가지게 되고, 이것의 TF-IDF 값도 커질 수 밖에 없다.

3) 실험 구현을 위해 데이터베이스관리시스템으로 Microsoft SQLServer를 사용하였으며, 개발 언어 및 미들웨어 JAVA와 JDBC(Java Database Connectivity)를 사용하였다.

〈표 3〉 정치 분야의 후보 키워드 리스트

순위	기존 TF-IDF	기존 TF-IDF의 수정		
		BTF	NTF1	NTF2
1	**일부**	이명박 후보	이명박 후보	**서울**
2	**박수**	후보	후보	**무단전재**
3	**협력**	**기자**	**기자**	**한국언론뉴스허브**
4	한나라당	에리카 김	에리카 김	**뉴시스통신사**
5	**서울**	**무단전재**	**무단 전재**	**재배포 금지**
6	**대표**	**광고**	**광고**	**모바일연합뉴스**
7	**오후**	**서울**	**서울**	**재배포금지**
8	**조성**	**재배포 금지**	**재배포 금지**	**저작권자연합뉴스**
9	**통합**	검찰	검찰	**오전**
10	**협상**	이명박	이명박	**오후**

수정 TF-IDF에 로그식의 적용(Normalized TFIDF)			
순위	BTF	NTF1	NTF2
1	이용득 위원장	이명박 후보	**백승렬**
2	이석행 위원장	에리카 김	**생각하십**
3	이회창 씨	검찰	**권주훈 기자**
4	금민 후보	BBK	**패널 질문**
5	**고 팀장**	김경준	대통합민주신당 정동영 대선후보
6	**망 설치 창원**	이명박	**이광호기자**
7	**준 뉴스앤조이**	**이전**	창당 10주년 행사
8	금액 조성	**기자회견**	**남강호기자**
9	**공기 음이온 보충**	**재배포 금지**	**박주성기자**
10	**초판 발행**	**서울**	**박지호**

주) 굵은 글씨체의 단어는 불용어 수준의 단어임

5.2 분야간 교차비교의 결과

2007년 11월 25일에 게재된 정치 분야 991개, 경제 분야 1025개, 연예 분야 480개 그리고 사회 분야의 뉴스 문서 1437개를 대상으로 분야간 교차비교를 실험하였다. 가장 기본적인 BTF와 TFIDF 식을 사용하여 각 10,000개의 단어를 가진 후보 키워드를 생성하였고 그 리스트들을 <표 4>에 제시하였다. 또한 이 키워드들에 분야간 교차비교분석을 적용하여 그 결과인 키워드들을 <표 5>에 제시하였다.

<표 5>의 각 분야 후보 키워드 집합에 굵

〈표 4〉 각 분야의 후보 키워드

순위	뉴스 분야			
	정치	경제	연예	사회
1	이명박 후보	**저작권자**	제28회 청룡영화상 시상식	**무단 전재**
2	김경준	**무단 전재**	영화	**뉴시스통신사**
3	검찰	**서울**	레드카펫	**재배포금지**
4	**한국언론 뉴스허브**	**재배포 금지**	해오름극장	**한국언론 뉴스허브**
5	**재배포금지**	**예정**	**재배포 금지**	**기자**
6	**뉴시스 통신사**	**현재**	청룡영화상	**서울**
7	**오후**	**무단전재**	보도자료	**오후**
8	**저작권자 연합뉴스**	**기자**	시상식	**저작권자 연합뉴스**
9	**오전**	**내년**	**저작권자**	**재배포금지**
10	**재배포 금지**	**리얼타임뉴스**	**중구**	**모바일 연합뉴스** 7070

주) 굵은 글씨체의 단어는 불용어 수준의 단어임

〈표 5〉 분야간 교차비교 후의 키워드 리스트

순위	뉴스 분야			
	정치	경제	연예	사회
1	이명박 후보	**리얼타임 뉴스**	제28회 청룡영화상 시상식	**검찰**
2	김경준	**아시아 경제**	영화	**방침**
3	검찰	**석간**	레드카펫	**지역**
4	**이명박**	**배포금지**	해오름극장	**협의**
5	**이면계약서**	**멀티미디어**	청룡영화상	**지역 빛**
6	**에리카 김**	**경제뉴스**	보도자료	**독자희망**
7	**계약서**	**기업**	시상식	**부산일보사**
8	**신당**	**외국인**	**장충동**	**이명박 후보**
9	**민주당**	**주가**	**국립극장**	**한나라당**
10	**김경준 씨**	**증시**	포즈	**김경준**

주) 굵은 글씨체의 단어는 교차비교를 이용, 불용어 수준의 단어를 제거한 후 상위 순위에 오른 단어임

게 표시된 단어인, '한국언론 뉴스허브', '재배포금지', '뉴시스 통신사' 등 의미 없는 단어가 많이 포함되어 있는 것을 확인할 수 있다. 하지만 분야간 교차비교를 통하여 <표 5>에서 보는 것처럼 의미 없는 단어가 제거되고 의미 있는 키워드가 추출되고 있음을 알 수 있다. 하지만 교차비교 분석을 통해 얻은 경제, 사회 분야의 키워드에는 몇 가지 불용어 수준의 단어가 포함되어 있는데, 이는 그 단어들이 여러 분야에 공통적으로 존재하지 않

은 경우이기 때문이다.

분야간 교차비교 시 사용되는 표준편차 임계값에 따른 키워드 리스트의 변화를 확인하는 것이 필요하다. 이 실험은 2008년 1월 3일에 게재된 정치 분야 1547개, 경제 분야 3587개, 연예 분야 1,232개, 사회 분야 803개의 뉴스가 포함된 분야별 문서 집합을 대상으로 하였다. 먼저 TF-IDF 변형식에 의하여 높은 가중치를 받은 상위 10,000개의 단어들로 분야별 후보키워드를 구성한다. 이에 대해 교차비교를 수행하여 의미없는 단어를 제거하게 되며, 이 때 적용하는 순위간 표준편차의 임계값을 10, 100, 1000, 10000으로 설정하였다 (<표 6> 참조). 여기서, 표준편차의 임계값을 10,000으로 설정한 것은 모든 분야의 후보 키워드집합에 등장한 단어는 제거됨을 의미한다.

<표 6>을 보면 표준편차의 범위가 커질수록 제거되는 단어가 많아짐을 알 수 있다. 하지만, 임계값을 10,000로 높였을 경우 의미있는 키워드까지 제거되는 현상이 발생하게 된다. 이로써 표준편차 임계값에 대한 적절한 설정을 통하여 키워드 선택이 가능함을 확인할 수 있었으며, 본 실험에서 사용한 뉴스 문서의 경우 적정한 표준편차 값은 100~1000 범위에 존재한다.

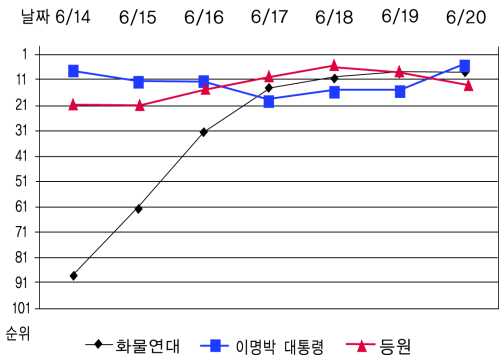
5.3 시간에 따른 키워드 집합의 추이 분석

본 절에서는 2008년 6월 14일부터 20일 기간 동안 정치 분야에 게재된 8,261개의 뉴스 문서들을 대상으로 키워드를 추출하여 시간에 따른 추이를 분석하였다. <그림 6>에서는 대상이 되는 7일 동안의 정치분야의 주요 쟁점과 관련하여 '화물연대', '이명박 대통령'과 '등원'의 세 키워드를 살펴보았다. '화물연대'의 경우, 6월 14일에는 88위에 불과했으나 6월 16일 이후에는 20위 내의 순위를 가졌다. 그리고 '이명박 대통령'과 '등원'이라는 단어는 계속 상위 순위에 있는 것을 확인할 수 있다. <그림 7>에서는 '쇠고기'를 포함한 7개

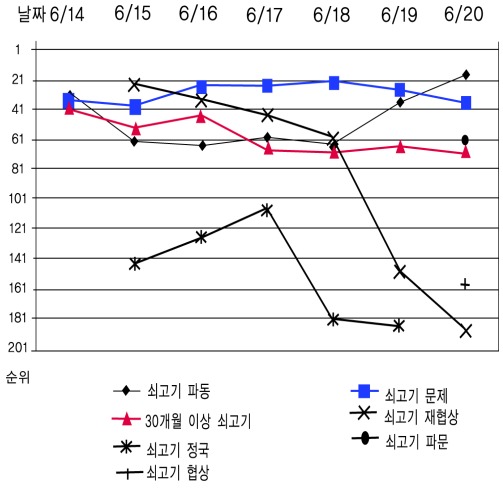
〈표 6〉 표준편차에 따른 분야별 교차비교 결과

순위	교차비교를 위한 표준 편차의 임계값			
	10	100	1000	10000
1	노 대통령	노 대통령	노 대통령	노 대통령
2	당선자	**당선자**	이명박 당선자	이명박 당선자
3	**광고**	이명박 당선자	인수위	인수위
4	이명박 당선자	**대통령**	**이명박 대통령 당선자**	노무현 대통령
5	대통령	**정부**	노무현 대통령	국회
6	**연합뉴스**	인수위	국회	총선
7	**저작권자 연합뉴스**	**이명박**	총선	인수위원회
8	정부	**대표**	인수위원회	대통령직
9	오전	이명박 대통령 당선자	대통령직	이명박 정부
10	인수위	**국민**	이명박 정부	정책

주) 굵게 표시한 단어는 표준편차 임계값의 증가에 따라 제거되는 단어임



〈그림 6〉 주요 키워드의 시간에 따른 추이



〈그림 7〉 '쇠고기'와 관련된 키워드의 시간에 따른 추이

키워드의 순위 변화를 관찰하였다. '쇠고기 문제'와 같은 단어는 꾸준히 상위 순위로 선정된 반면, '쇠고기 재협상'과 같은 단어는 처음에는 상위 순위로 선정되었으나 시간이 지나면서 순위가 하락하는 것을 확인할 수 있다.

<그림 6>, <그림 7>의 분석을 통해, 시간의 추이에 따라 키워드 순위가 사회적 이슈 및 사건과 동조함을 알 수 있었다. 즉 사회적 이슈를 대표하는 관련 키워드가 상위 순위로 올라오고, 그렇지 않은 키워드는 시간이 지남에 따라 관심도가 저하되는 현상을 파악할 수 있다. 이는 제안 기법이 뉴스문서집합에서 키워드를 효과적으로 추출할 수 있음을 보여주는 것이다.

6. 결론 및 향후 연구

본 논문에서는 인터넷 포털 사이트에 게재되는 대용량 뉴스기사로부터 요약하여 보여줄 수 있는 방법으로 키워드 추출을 제시하였다. 분야별 뉴스에서 의미 있는 키워드를 추출하기 위해서, 검색엔진 개발에 활용되고 있는 기존의 TF-IDF 가중치 모델을 변형하여, 전체 문서집합에 적용할 수 있는 6가지 TF-IDF 변형식을 제안하였으며, 분야간 교차비교 분석을 통해 불용어 수준의 키워드를 제거하였다. 제안 기법의 효용성을 검증하기 위해 국내의 대표적 인터넷 포털 사이트인 네이버에 게재되는 뉴스문서에서 추출한 키워드의 품질을 평가하였으며, 이러한 키워드들이 실제로 사회적 관심도의 변화에 따라 그 순위가 변화하는지를 관찰하였다.

본 논문에서 제안한 키워드 추출 기법은 다른 종류의 문서 집합의 주제 탐색에도 활용될 수 있으며, 특히 우리는 검색엔진의 개인화와 오피니언 마이닝(Opinion Mining)의 연구에 활용할 예정이다. 검색엔진의 개인화는 사용자 성향을 반영한 검색어의 확장으로 구현될 수 있는데, 여기서 본 논문의 제안 기법을 적용하여 사용자가 과거 선택한 웹페이지에서 주요 키워드를 추출, 사용자 프로파일을 구성할 수 있다. 오피니언 마이닝은 상품평 등의 오피니언 데이터로부터 사용자의 주

관적인 의견 정보를 추출하는 분야로서, 핵심적인 연구 이슈가 주관적 의지를 담고 있는 감정단어와 이에 관련된 속성을 찾는 것이다. 본 제안 기법은 데이터의 특성을 감안한 튜닝 및 기계 학습 알고리즘의 융합을 통해 오피니언 속성 추출 기법으로 승화시킬 것이다.

참 고 문 헌

[1] 강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학출판사, 서울, 2002, pp. 507-549.

[2] 한국인터넷진흥원, "2007년 하반기 정보화 실태조사 요약 보고서", 2008.

[3] Jo, Taeho, Lee, Malrey, and Gatton, T. M, "Keyword extraction from documents using a neural network model," ICHIT'06, Vol. 2, 2006, pp. 194-197.

[4] Matsuo, Y., and Ishizuka, M., "Keyword extraction from a single document using word co-occurrence statistical information," International Journal on Artificial Intelligence Tools, Vol. 13, No. 1, 2003, pp. 157-169.

[5] Robertson, S., "Understanding inverse document frequency : on theoretical arguments for IDF," Journal of Documentation, Vol. 60, No. 5, 2004, pp. 503-520.

[6] Robertson, S. E., "Term specificity," Journal of Documentation, Vol. 28, 1972, pp. 164-165.

[7] Robertson, S. E., "Specificity and weighted retrieval," Journal of Documentation, Vol. 30, No. 1, 1974, pp. 41-46.

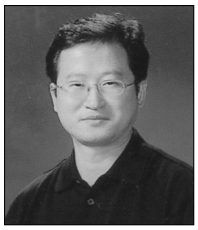
[8] Robertson, S. E., "The probability ranking principle in information retrieval," Journal of Documentation, Vol. 33, 1977, pp. 294-304.

[9] Wang, J., Liu, J., Wang, and Cong, "Keyword extraction based on PageRank," Lecture notes in computer science, 2007, pp. 857-864.

[10] Yu, J. X., Kitsuregawa, M., and Leong, H. V., "Keyword Extraction using Support Vector Machine," Lecture notes in computer science, Vol. 4016, 2006, pp. 85-96.

저 자 소 개

이성직 (E-mail : sungjick@gmail.com)
2008년 서울시립대학교 전자전기컴퓨터공학부 졸업 (공학사)
2008년~현재 서울시립대학교 전자전기컴퓨터공학부 석사과정
관심분야 정보검색, 데이터베이스



김한준 (E-mail : khj@uos.ac.kr)
1994년 서울대학교 계산통계학과 졸업 (이학사)
1996년 서울대학교 전산과학과 대학원 졸업 (이학석사)
2002년 서울대학교 컴퓨터공학부 대학원 졸업 (공학박사)
2002년~2002년 서울대학교 공과대학 Post-Doc
2002년~현재 서울시립대학교 전자전기컴퓨터공학부 부교수
관심분야 텍스트마이닝, e-비즈니스 기술, 정보검색, 데이터베이스