주제 추적을 위한 뉴스 키워드 추출 기법

이성직° 김한준
서울시립대학교 전자전기컴퓨터공학부
sjleekor@gmail.com, khj@uos.ac.kr

News Keyword Extraction for Topic Tracking

Sungjick Lee° Han-joon Kim
Department of Electrical and Computer Engineering, University of Seoul, Korea

본 논문에서는 인터넷 포털사이트에 게재되는 대용량 분야별 뉴스문서집합을 대상으로 키워드를 추출하여 분야별 주제를 제시할 수 있는 키워드추출 기법을 제안한다. 키워드추출은 정보검색, 문서분류, 주제탐색, 문서요약 등의 다양한 텍스트마이닝 분야의 연구에서 주요 속성을 추출하기 위해 이용되고 있어, 그 활용 가치가 크다. 본 논문은 인터넷 포털사이트 네이버에 게재되는 하루 10000개 이상의 뉴스문서를 대상으로 키워드를 추출하고자 한다. 이러한 대용량의 뉴스문서집합에서 키워드를 추출하여 제시하고 이 키워드가 추출된 뉴스 기사를 보여줄 수 있다면 이용자들은 화제가 되고 있는 뉴스 문서를 효율적으로 검색할 수 있을 것이다. 이를 위해, TF-IDF(Term Frequency-Inverse Document Frequency) 가중치를 변형하여 각 분야에서의 키워드를 추출할 수 있는 자동화된 시스템을 고안하였다. 시스템은, 우선 변형된 TF-IDF 가중치를 이용하여 기사에 등장한 단어들을 정치, 경제, 사회, 연예 등의 분야별로 순위를 선정하여 '후보 키워드리스트'를 얻는다. 그리고 이 리스트에서 중요하지 않다고 생각되는 단어를 제거하기 위해 분야간 교차비교 기법을 사용하여 최종적으로 '키워드리스트'를 얻는다.

대용량의 분야별 뉴스문서집합에서 키워드를 추출하기 위해 먼저 대상 인터넷사이트에서 뉴스문서를 읽어 와서 데이터베이스에 저장한다. 저장된 뉴스문서에 대하여 등장한 명사를 대상으로 변형된 TF-IDF 가중치를 이용하여 큰 가중치를 얻은 단어들로 '후보 키워드리스트'를 구성한다. 기본적으로 TF-IDF 모델은 어떠한 문서집합이 있을 때, 특정 문서 내에서의 단어들 간의 중요성을 판단하는 척도로서 이용된다. 이 TF-IDF 값은 일반적으로 TF 값과 IDF 값의 곱으로 계산된다. 이 때, TF 값은 문서 내에서 많이 발생한 단어가 중요하다는 의미를 갖고, IDF 값은 발생한 문서의 수가 적은 단어일수록 더 중요하다는 의미로 사용된다.

본 논문에서는 특정 문서에서의 중요성이 아닌 문서집합에서 중요하다고 생각되는 단어를 찾는 것이 목적이므로 원래의 모델을 변형하여 사용한다. 그 변형된 모델을 그림 1과 같다.

Term Frequency			Inverse Document Frequency
$\sum_{j=1}^{\lvert D\rvert} n_{i,j}$	$\frac{TF_i}{Max(TF_1, TF_2, \ldots, TF_{\lvert T\rvert})}$	$\sum_{j=1}^{\lvert D\rvert} \frac{n_{i,j}}{\sum_{k=1}^{\lvert T_j\rvert} n_{k,j}}$	$\log \frac{\lvert D\rvert}{\lvert\{d_j : t_i \in d_j\}\rvert}$

TF-IDF	
$TF_i \times IDF_i$	$(TF_i + 1.0) \times IDF_i$

그림 1. TF, IDF, TFIDF 변형식

먼저 TF 값을 계산하기 위해 하나의 기본식과 두 개의 표준화 식을 사용한다. 먼저 기본식은 특정 단어가 각 문서에서 출현한 회수를 단순히 더한 값을 계산한다. 그런데 단순한 출현빈도의 합은 단어 간의 상대적 중요도를 표한하기 어렵기 때문에 이를 표준화한 다음의 두 가지 식을 추가로 사용한다. 첫 번째 표준화 식은 기본식으로 계산된 값 중에서 가장 큰 값으로 나누는 방법을 사용한다. 두 번째 표준

화 식은 더 긴 뉴스 기사에서 등장하고 있는 단어가 상대적으로 큰 TF 값을 갖지 않도록 하기 위해서 사용한다. 이 식은 각 문서에서의 해당 단어의 발생빈도를 모든 단어의 발생빈도로 나눈 다음 그 값들을 더하여 TF값을 구한다. 그리고 IDF 값은 일반적으로 사용되는, 총 문서의 수를 특정 단어가 출현한 문서의 수로 나누어 로그를 취하여 계산한다. 마지막으로 TF-IDF 값은 기본적으로 TF 값과 IDF 값을 곱하여 얻는다. 그리고 로그를 취하여 얻는 IDF에 비해 TF 값이 매우 커질 수 있기 때문에 TF 값에 로그를 취하고, 1을 더하여 IDF 값과 곱하는 변형식을 추가로 사용한다. 위의 식을 계산하여 얻은 각 단어에 대한 TF-IDF에 대한 가중치를 이용하여 순위가 매겨진 분야별 '후보 키워드리스트'를 얻는다.

본 논문에서는 분야별 뉴스문서집합에서 각각 키워드를 얻고자 한다. 그런데 각각의 분야별 뉴스문서집합은 '일반적인' 뉴스문서집합의 부분집합이다. 따라서 분야별 뉴스 문서집합에 대하여 변형된 TF-IDF 가중치를 매길 경우, 일반적인 뉴스 문서집합 전체에서 중요한 단어들도 큰 가중치를 얻을 수 있다. 이런 단어들은 각 분야의 뉴스 문서집합에서는 키워드로서 부적절하므로 제거하는 것이 바람직하다. 이를 위해 본 논문에서는 분야간 교차비교를 이용하여 제거할 단어를 선정한다.

제거할 단어의 선정은 기본적으로 모든 분야의 '후보 키워드리스트'에서 일정 순위 이상을 받은 단어를 대상으로 한다. 본 논문에서는 각 분야의 '후보 키워드리스트'의 상위 5000 개의 단어를 대상으로 분야간 교차비교를 하였다. 이러한 교차비교는 키워드 순위에 대한 표준편차 값을 계산함으로써 이루어진다. 표준편차 값을 계산하여 일정 임계값 이하이면 제거할 단어로 선정한다. 순위간 표준편차를 이용해 제거할 단어를 한정하는 이유는 특정 단어의 각 분야에서의 순위가 크게 차이가 날 경우, 상위 순위에 있는 분야에서는 해당 단어를 키워드로 볼 수 있기 때문이다.

이와 같이 '후보 키워드리스트'의 순위를 이용하여 '일반적인' 뉴스문서집합의 키워드를 제거하고, 최종적으로 각 분야별 뉴스문서집합의 '키워드리스트'를 얻는다.

본 논문에서 제안하고 있는 키워드추출 기법을 실험하기 위해 인터넷 포털사이트 네이버에 게재되고 있는 뉴스기사를 대상으로 하였다. 이 사이트에서 분야별로 나누어 제공하고 있는 각 언론사의 뉴스문서 중에서 정치, 경제, 연예, 사회 분야의 뉴스문서를 수집하여 키워드를 추출하였다. 먼저 뉴스문서를 데이터베이스에 저장한 후에 이 뉴스 문서들에 대하여 단일명사를 추출하고, 복합명사를 합성한다. 그리고 각 문서에서 단일명사와 복합명사가 출현한 사실을 데이터 웨어하우스(Data Warehouse)에서 사용하는 스타스키마(Star Schema)와 같은 형태로 저장하였다. 이 출현사실을 활용하여 변형된 TF-IDF 가중치를 계산하였고, '후보 키워드리스트'를 얻었다. 그리고 최종적으로 분야간 교차비교를 통해 '키워드리스트'를 얻었다.

표 1. 추출된 키워드 리스트 (좋은 결과는 진하게 표시)

Rank	뉴스문서의 분야			
	정치	경제	연예	사회
1	**이명박 후보**	리얼타임 뉴스	**제28회 청룡영화상 시상식**	검찰
2	**김경준**	아시아 경제	**영화**	방침
3	**김찬**	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**	**증시**	**포즈**	**김경준**

표 1은 2007년 11월 25일에 게재된 뉴스문서를 대상으로 추출한 키워드이다. 변형된 TF-IDF 가중치를 이용한 실험에서 정치, 연예 분야의 뉴스문서집합에서는 의미 있는 키워드를 추출할 수 있었으나, 경제, 사회 분야의 뉴스문서집합에서는 비교적 좋지 못한 결과를 얻었다. 이러한 결과를 개선할 수 있도록 보다 효과적으로 불용어 수준의 단어를 제거할 수 있는 방법을 고안하고 있다. 또한 뉴스문서집합의 각 분야에 대해서 주제 추적을 할 수 있는 시스템을 개발하고 있다.