단어 분별도에 기반한 뉴스 검색 문서 요약

이상건○, 이혜민, 김기령, 서덕호, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
{lsk3020, hyemm1215, kim005kim}@naver.com, sdh4389@nate.com, halee@kumoh.ac.kr

Search Resulted News Summarization using Word Discriminability

Sang-Keon Lee○, Hye-Min Lee, Gi-Ryeong Kim, Duc-Ho Seo, Hyun Ah Lee
Dept. of Software Engineering, Kumoh National Institute of Technology.

요 약

다양한 언론사로부터 기사를 제공받아 서비스하는 인터넷 포털의 뉴스에서는 수많은 중복 기사가 실시간으로 등록된다. 이로 인하여 인터넷 포털에서 관심 있는 주제의 기사를 검색하여 찾아보려는 경우 검색 키워드를 포함한 기사의 수가 지나치게 많아 원하는 정보를 적절하게 얻기 어렵다. 본 논문에서는 이러한 문제점을 해결하기 위해서 검색 기사 중 유사한 문서를 군집화하고 군집에 대한 다중문서요약을 사용자에게 제시하여 검색된 기사를 효율적으로 활용할 수 있는 방법을 제시한다. 다중문서 요약에서는 뉴스 기사에 적합한 단어 가중치인 분별도(discriminability)를 제안하여 사용하여 군집화된 기사로부터 유사 문장을 군집한다. 시스템에서는 군집된 기사의 대표 문장 군집에서 대표 문장, 즉 키워드에 대한 주제별 기사의 요약문을 결과로 제시하여, 효율적인 뉴스 검색을 지원한다.

주제어: 분별도, 뉴스 검색, 뉴스 문서 요약, 다중 문서 요약

1. 서론

인터넷의 급속한 성장과 그에 따른 데이터양의 증가로 생활에 필요한 모든 정보를 인터넷에서 얻을 수 있다. 그러나 하루에도 엄청나게 많은 데이터가 발생하기 때문에 사용자가 이 모든 데이터를 확인하는 것은 현실적으로 불가능하다. 특히 인터넷 뉴스는 실시간으로 수많은 중복된 기사들이 발생하여 원하는 정보를 찾아보기 쉽지 않다. 이러한 문제를 해결하기 위해 개인화 정보를 사용하여 사용자에게 적합한 뉴스를 제공하거나[1][2], 연관 뉴스를 제시하는 방법[3] 등이 제시되었으나, 이러한 방식은 사용자의 유동적인 관심 이슈를 반영할 수 없고, 정확하지 않은 개인화 정보로 인한 문제를 가진다.

관심 있는 이슈에 대한 기사를 찾아보기 위해서 대부분의 사용자는 키워드 검색을 사용한다. 하지만 발생되는 기사가 많고 최근 기사를 우선적으로 보여주는 뉴스 검색에서는 시간이 지난 기사에 대한 정보를 찾아보기 쉽지 않아 정확한 정보 획득이 용이하지 않다. 또한 같은 이슈에 대하여 각기 다른 제목을 사용하여 여러 언론사에서 제공되는 뉴스의 경우에는 사용자가 직접 뉴스 기사를 읽어보지 않으면 정확한 내용을 확인하기 어렵다. 이와 같이 뉴스 검색의 경우 검색된 결과에서도 중복을 포함한 수많은 뉴스 기사가 제시되어 관심 있는 이슈에 대한 사용자의 정보 요구를 만족시키지 못한다.

본 논문에서는 이러한 문제를 해결하기 위해서 키워드에 의해 검색된 뉴스 기사에 대한 다중문서요약을 제안한다. 기존 다중문서 요약에서는 동일 주제의 문장에서 요약을 생성하는 것[4][5]을 문제로 한다. 이와는 다르게 제안하는 시스템에서는 키워드로 검색된 뉴스를 대상으로 하고 있으므로, 시스템에서는 첫 단계로 검색된 뉴스를 수집하고, 다음 단계에서 유사한 뉴스 기사들의 군집화 과정을 수행한다. 군집화된 기사에 대한 다중문서 요약을 수행하여 유사한 기사들의 대표 문장 즉 요약을 제시하면, 중복된 기사에 의한 문제점을 해결할 수 있을 뿐만 아니라 시간의 흐름에 따른 뉴스의 흐름도 쉽게 파악할 수 있다.

유사 기사의 요약인 기사 군집의 대표 문장을 추출하기 위해서는 군집 내의 유사한 문장을 다시 군집화하고 군집 내의 대표 문장을 추출해야 한다. 문장 군집화에서는 여러 문장에서 자주 사용되는 단어보다 특정 문장에서 집중적으로 사용하는 단어의 가중치가 높게 측정되어야 한다. 본 논문에서는 이러한 단어 가중치를 얻기 위해 단어 분별도(discriminability)를 제안하여 사용한다. 단어 분별도는 조건부 확률에 기반하여 계산한다. 시스템에서는 분별도를 이용하여 문장 군집을 수행한 뒤 군집 크기로 정렬하여 대표 문장을 추출하여 사용자에게 제시한다.

그림 1은 제안하는 시스템의 구조를 보인다. 시스템은 문서 수집과 군집, 문장 군집과 요약문 추출로 구성된다. 문서 수집에서는 다양한 언론사의 뉴스를 제공하는 네이버 뉴스 검색 API를 통해 최신 기사를 수집한다. 다음에서는 문서 군집에서 요약문 추출까지의 단계에 대해서 상세히 설명한다.

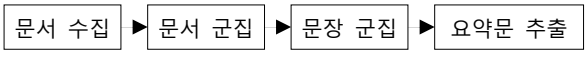

그림 1 시스템 구조

2. 단어 분별도를 기반한 뉴스 검색 문서 요약

뉴스 검색 문서에 대한 요약은 수집된 문서에 대한 문서 군집, 문장 군집, 요약문 추출으로 처리된다. 본 논

문에서는 뉴스 기사의 특성을 고려한 비교적 간단한 군집화 방식을 적용한 뒤, 분별도를 이용하여 문장 군집을 수행한다. 아래에서 각 단계에 대해 설명한다.

2.1 검색 기사 군집화

키워드로 검색된 최신 뉴스 문서 집합에는 동일한 내용에 대한 다른 각기 다른 언론사의 기사가 중복해서 등장한다. '황산'과 같이 중의적(예를 들어 '황산 테러'와 '중국 황산 자매 결연')인 키워드이거나, 그림 2의 '변호사'와 같이 사회면이나 연예면 등의 여러 분류에서 등장하는 키워드의 경우에는 각기 다른 주제의 기사들이 검색 결과에 혼합되어 나타나기도 한다. 또한, 그림 3의 '아시안게임'에 대한 기사가 2014년 9월 28일에는 야구에 대한 내용이었다가 이후 30일에는 축구나 다른 종목에 대한 내용으로 바뀌는 것처럼, 키워드에 대해 새로운 이슈가 발생하면 검색된 문서들의 주제가 바뀌기도 한다. 이처럼 뉴스 기사의 경우 하나의 키워드에 대해 검색된 기사들은 다양한 주제를 다루고 있으므로, 뉴스 검색 결과의 사용성을 높이기 위해서는 유사한 주제를 가진 문서 군을 생성하기 위하여 문서 분류 작업을 수행해야 한다.

문서군을 생성하는 방법은 코사인 유사도를 이용한다. 문서 분류를 위한 문서 특성은 제목과 본문에서 사용되는 명사를 이용하여 구성한다. 형태소 분석기를 이용하여 문서 내의 단어를 추출한 뒤, 각 단어의 빈도를 얻는다. 코사인 유사도에서 각 문서의 벡터를 구하기 위한 단어 가중치는 문서에서 발생한 단어의 빈도 tf에 단어의 idf값을 곱하여 얻는다. 각 문서의 단어 점수를 이용하여 벡터를 구성하고, 각 문서의 벡터들을 이용하여 코사인 유사도 값을 구하여 유사한 주제의 뉴스 문서 군집을 구성한다. 군집화에서는 비교적 간단한 알고리즘을 적용한다. 최신기사 우선으로 기사 검색 결과를 얻고, 입력된 기사 순서대로 기사의 유사도를 계산하여 군집화를 수행한다. 기사는 기사에 포함되어 있는 단어들로 벡터를 구성하고, 군집은 군집에 포함되어 있는 단어들로 벡터를 구성한다. 기사와 군집 간 코사인 유사도가 임계치를 넘으면 유사한 주제의 문서로 판단을 하여 군집에 포함시킨다. 임계치를 넘는 기사가 존재하지 않으면 새로운 군집을 구성한다. 뉴스 기사는 시간의 흐름에 따라 유사한 기사들이 동시 다발적으로 발생하는 특성이 있어 이와 같은 간단한 군집화 방식으로도 비교적 좋은 성능을 얻을 수 있다. 본 눈문에서는 코사인 유사도의 임계치로 실험적으로 얻은 0.3을 사용한다.

그림 2. '변호사' 뉴스 검색 결과

그림 3. 2014/9/28 '아시안게임' 뉴스 검색 결과

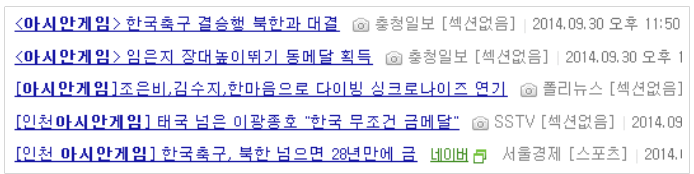

그림 4. 2014/9/30 '아시안게임' 뉴스 검색 결과

2.2 분별도를 이용한 문장 군집화

유사한 주제를 가진 각각의 문서 군집에서 요약 문장을 추출하기 위해서 군집 내에 있는 문서들을 문장 단위로 분리하고 유사 문장을 군집화를 하는 과정을 거친다.

문장 군집화를 위한 문장 벡터에는 tf-idf를 이용한 가중치 계산은 적절하지 않다. 문장 군집화에 참여하는 문서 집합은 키워드에 검색된 기사들을 군집화하여 얻는 문서이므로, 각 문서가 가지고 있는 단어들 또한 유사하다. 이 경우 문서 집합 내의 특정 문서에서만 등장하는 단어에 높은 가중치를 주는 idf는 적절한 가중치를 제공하지 않는다[6]. 따라서 문장 군집화를 위해서는 별도의 단어 가중치 기법이 필요하다.

많은 정보를 한 문장에 모두 표현할 수도 있고, 여러 문장에 나누어 표현할 수도 있다. 그림 5는 키워드에 대해 수집된 뉴스 문서 집합에서 등장한 문장들과 문장에 포함된 단어들의 한 예를 보인다. 그림에서의 단어 분포를 보면 문장 $\{S_1, S_2, S_3, S_4\}$, $\{S_5, S_6, S_7\}$, $\{S_8, S_9, S_{10}\}$이 각각 군집을 구성할 수 있다. 이 때, 단어 w_1과 w_4는 경우 모든 군집에 걸쳐 등장하고, 단어 w_3, w_5, w_6, w_8, w_{10}은 특정 문장 군집에서만 등장한다.

문장의 유사도를 판단하기 위해서는 각 문장에서 다른 문장들과 차이를 표현해 주는 단어를 찾을 수 있어야 한다. w_1과 w_4는 거의 모든 문장에 등장한다. 이러한 단어들은 다른 문장과의 차별성을 표시해 주지 못한다(이건 idf에서도 낮은 값으로 나옴). 반면, 단어 w_3, w_5, w_6, w_8, w_{10}은 특정 군집에만 등장한다. 예를 들어 단어 w_3은 다른 군집에는 등장하지 않는 단어이므로 $\{S_8, S_9, S_{10}\}$만의 차별성을 나타낸다고 볼 수 있다. 본 논문에서는 조건부 확률에 기반한 분별도를 이용하여 군집에서의 단어 차별성을 측정한다. 단어 w에 대한 단어 w_i의 조건

단어 / 문장	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}
S_1	O	O						O		O
S_2	O	O		O				O	O	O
S_3	O			O				O	O	O
S_4	O						O	O		O
S_5	O	O		O		O	O			
S_6	O			O		O	O			
S_7	O	O				O				
S_8	O	O	O	O	O				O	
S_9	O		O	O	O					
S_{10}	O		O	O	O					

그림 5. 간단한 문장 모델

부 확률 $P(w_i|w)$는 아래 식(1)을 이용해서 구할 수 있다. 단어 w와 단어 w_i가 동시에 등장한 문장의 개수를 단어 w가 등장한 문장의 개수로 나누어 값을 구한다. 조건부 확률은 한번이라도 한 문장에서 같이 등장한 적이 있는 단어들에 대해서만 구한다.

$$P(w_i|w) = \frac{count_s(w, w_i)}{count_s(w)} \quad (1)$$

아래 그림 6은 그림 5의 w_1과 w_3, w_9에 대한 조건부 확률을 구한다. 단어 w_1은 10개의 문장에서 등장하였으며, 다른 모든 단어들과 공기한다. 따라서 나머지 모든 단어와 각각 조건부 확률을 구할 수 있다. 단어 w_3은 세 문장에서만 등장하였고, 단어 w_1, w_2, w_4, w_5, w_9와만 공기하므로, 이 단어들로부터 조건부 확률을 구한다. 단어 w_9도 세 문장에서만 등장하였고, 공기한 단어 w_1, w_2, w_3, w_4, w_5, w_8, w_{10}로부터 조건부 확률을 구한다. 그 결과로 w_1보다 w_3의 조건부 확률의 평균이 높게 나타나, w_3이 문장의 차이성을 표현해 주는 단어임을 알 수 있다.

단어	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	평균
w_1		$\frac{5}{10}$	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	0.39
w_3	$\frac{3}{3}$	$\frac{1}{3}$		$\frac{3}{3}$	$\frac{3}{3}$				$\frac{1}{3}$		0.73
w_9	$\frac{3}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{3}{3}$	$\frac{1}{3}$			$\frac{2}{3}$		$\frac{2}{3}$	0.67

그림 6. 단어 1, 단어 3에 대한 조건부 확률

단어의 문장 차이성을 표현해 주는 값을 본 논문에서는 분별도(discriminability)로 표현한다. 분별도는 식(1)의 조건부 확률에 기반하여 식 (2)로 구한다.

$$Discriminability(w) = \frac{\Sigma_{w_i \in CO_w} P(w_i|w)}{|CO_w|} \quad (2)$$

식에서 CO_w는 단어 w와 같이 등장한 적이 있는 단어들의 집합을 나타낸다. 예를 들어 CO_{w_1}은 {S_2, S_8, S_{10}}이 된다. 단어 w에 대해 구한 각 공기 단어의 조건부 확률을 모두 더하고, 집합 CO_w의 크기 $|CO_w|$로 나누어 평균 분별도를 구한다. 예에서 단어 w_1과 단어 w_3의 가중치를 계산하면 단어 w_1의 가중치로 0.39, 단어 w_3의 가중치로 0.73, 단어 w_9의 가중치로 0.67을 얻는다. 결과에서 동일한 idf를 가진 w_3과 w_9중에서 특정 군집에서 집중적으로 나타나는 w_3가 더 높은 가중치를 가지는 결과를 볼 수 있다. 이와 같은 분별도는 참여한 문장과 단어의 개수가 많아지면 값의 차이가 두드러지는 결과를 얻을 수 있다.

얻어진 분별도를 단어 가중치로 사용하여 문장에 대한 단어 벡터를 구하고, 코사인 유사도를 이용하여 문장 유사도를 계산한다. 얻어진 유사도 값을 이용하여 문서 군집화와 동일한 방식을 적용하여 군집화를 수행한다. 문장 군집화에서도 실험에서 얻은 평균적인 성능을 보이는 임계치 0.3을 사용한다.

2.3 요약 정보 추출

최종적으로 뉴스 검색 결과에 대한 다중 문서 요약의 결과로 군집화된 기사의 요약문인 대표 문장을 추출한다. 군집화된 기사에 대한 문장 군집의 결과로 여러 개의 군집을 얻을 수 있는데, 각 군집의 크기가 클수록 즉 군집에 포함된 문장이 많을수록 여러 기사에서 자주 발생하는 중요 문장으로 볼 수 있다. 시스템에서는 문장 군집의 크기의 내림차순으로 정렬하고, 상위 문장 군집의 대표 문장을 요약 정보로 제시한다. 군집의 대표 문장을 추출하기 위해서는 가장 많은 명사를 갖는 문장이 가장 많은 정보를 담고 있는 문장이라 판단하여 문장 군집의 대표 문장으로 선택한다.

3. 실험

실험을 위해 2014년 9월 이슈화된 키워드 '담배값'에 대한 뉴스 문서를 수집하여 평가를 수행하였다.

그림 7은 '담배값'에 대한 검색 문서를 군집화한 뒤 한 군집에서 나타나는 문장 예를 보인다. 그림의 군집에서는 담뱃값에 대한 다양한 기사 중 '임종규 국장'의 발언 중심으로 군집화된 것을 알 수 있다.

표 1은 그림 4의 군집에서 발생하는 단어들을 빈도 내림차순으로 보이고, 이에 대한 분별도의 순위는 괄호로

	문장
1	임 국장은 구체적인 담배값 가격 인상 폭과 관련해서 "아직 논의가 필요한 부분이지만 '상당 폭'을 올려야 효과가 있을 것"이라고 말했다.
2	임 국장은 "구체적인 담배값 가격 인상 폭과 관련해서 아직 논의가 필요한 부분이지만 상당폭을 올려야 효과가 있을 것"이라고 전했다.
3	지난 11일 임종규 보건복지부 건강정책 국장은 "복지부로서는 세계보건기구(WHO)의 담뱃세 인상 권고를 받아들여 담배규제기본협약(FCTC) 당사국으로서 담뱃세 인상을 강하게 추진할 계획"이라고 밝혔다.
4	11일 임종규 보건복지부 건강정책 국장은 "복지부로서는 세계보건기구(WHO)의 담뱃세 인상 권고를 받아들여 담배규제기본협약(FCTC) 당사국으로서 담뱃세 인상을 강하게 추진할 계획"이라고 말했다.

그림 7. 문장 군집에 속한 문장들

단어	빈도	분별도	단어	빈도	분별도
인상	75	0.006432 (18)	계획	36	0.050636 (8)
보건	72	0.023983 (14)	권고	36	0.102302 (3)
담뱃세	66	0.027203 (12)	기구	36	0.064458 (6)
국장	49	0.202341 (2)	복지	36	0.031784 (10)
임	49	0.144183 (4)	부	36	0.045245 (9)
담배	47	0.006559 (17)	정책	36	0.029038 (11)
종규	40	0.292551 (1)	추진	36	0.024243 (13)
세계	37	0.059962 (6)	규제	35	0.094337 (5)
건강	36	0.021807 (15)			

표 1. 문장 군집 단어, 빈도, 분별도

표시한다. 표에서 볼 수 있듯이 '인상', '보건', '담뱃세', '담배'는 '담배값'으로 검색한 9월의 대부분의 기사의 문장에 포함되어 높은 빈도를 가지지만, 제안한 분별도는 낮은 값을 가진다. 이에 비하여 군집에서 특징적으로 발생하는 '종규', '국장'과 같은 단어의 분별도 값은 비교적 높게 나온 것을 볼 수 있다.

그림 8은 '담배값'에 대한 2014년 9월 19일 기준의 시스템 결과를 보인다. 결과에서 볼 수 있듯이 키워드에 대한 검색 문서 중 유사한 문서를 군집화하고 군집을 대표하는 요약 문장을 선택하여, 기존 뉴스 검색이 가지는 최신 기사 중심의 결과 제공의 단점을 해소하여, 키워드에 관련된 다양한 결과를 한 눈에 볼 수 있다.

1	정부는 담배에 붙는 세금에 지방세인 '안전세'를 신설해 담뱃값을 인상하는 방안을 검토 중인 것으로 알려졌다.
2	지난 3일 보건복지부와 한국건강증진개발원이 전국 만 19세 이상 성인 1천 명을 대상으로 전화 설문조사를 실시한 결과 담뱃값이 4,500원으로 인상될 경우 담배를 끊겠다는 응답이 32.2%로 나타났다.
3	현재 성인 흡연율은 세계에서 가장 높은 수준인데 방치하고 있는 것이나 다름 없다면서, 이번 기회에 흡연률을 선진국 수준으로 낮출 수 있도록 금연 예방 대책을 추진하고, 국민들도 정부의 노력에 이해해달라고 말했습니다.
4	11일 임종규 보건복지부 건강정책 국장은 "복지부로서는 세계보건기구(WHO)의 담뱃세 인상 권고를 받아들여 담배규제기본협약(FCTC) 당사국으로서 담뱃세 인상을 강하게 추진할 계획"이라고 말했다.
5	기획재정부는 담배가격 인상안 확정 발표 후 담배 판매량 급증과 품귀현상이 예상됨에 따라 담배시장 질서 교란 방지를 위해 '매점매석 행위에 대한 고시'를 이날 정오부터 담뱃값이 인상되는 날까지 한시적으로 시행한다고 밝혔습니다.

그림 8. 키워드 '담배값'에 대한 요약문 추출 결과

4. 결론

본 논문에서는 뉴스에 대한 키워드 검색 결과 문서의 특성을 고려하여 문서 군집화와 문장 군집화를 수행한 뒤 대표 문장을 선정하여 뉴스 검색 결과를 효율적으로 사용하기 위한 방법을 제안하였다. 시스템에서는 동일한 주제에 대하여 수집된 최신 뉴스 기사 집합들은 이미 유사한 단어들을 갖고 있으므로 tf-idf를 활용하는 방법이 적절하지 않다. 이를 고려하여 본 논문에서는 조건부 확률에 기반한 분별도로 단어의 가중치를 표현하여, 코사인 유사도를 활용한 유사 문장 군집화를 하는 방법을 제시하였다. 실험 결과 특정 문장에만 등장하는 차별성이 있는 단어들에 의해 문장들이 군집화되는 결과를 얻을 수 있었다.

향후 연구로는 다양한 실험을 통한 분별도의 개선과 코사인 유사도 이외의 유사도 계산 방식의 적용, 추가적인 군집화 알고리즘의 적용 등을 예정하고 있다.

참고문헌

[1] Zhu, J., Wang, C., He, X., Bu, J., Chen, C., Shang, S., Qu, M., Lu, G., "Tag-oriented document summarization," Proceedings of 18th International World Wide Web Conference (WWW`09), 195-1196, 2009

[2] 김태환 "개인화 정보를 이용한 자동 뉴스 피딩 시스템", 한양대학교 컴퓨터공학과 박사 학위 논문, 2012.

[3] 김지혜, 장재영, 윤홍준, 김한준, "키워드 관련도를 이용한 뉴스기사의 연관검색 기법", 한국컴퓨터종합학술대회 논문집, 37(1C), 53-57, 2010.

[4] Inderjeet Mani, “Automatic Summarization”, Kohn Benjamins Publishing Co., 2001.

[5] 이일주, 김민구, "단어의 공기정보를 이용한 클러스터 기반 다중문서 요약", 정보과학회논문지 : 소프트웨어 및 응용, 33(2), 243-251, 2006.

[6] 정석팔, 임성현, 전진형, 김병만, 이현아, “요약문을 이용한 웹 검색 결과 군집화”, 정보과학회논문지:데이타베이스, 39(5), 321-331, 2012.