韓國컴퓨터情報學會 論文誌
第15卷 第11號, 2010. 11.

2010-15-11-1-3

의견 문서의 단어 통계 분석을 통한 의견 검색 특성에 관한 연구

한 경 수*

A Study on the Characteristics of Opinion Retrieval Using Term Statistical Analysis in Opinion Documents

Kyoung-Soo Han*

요 약

문서에 표출된 사용자의 의견을 검색하는 의견 검색의 성능이 일반 사실을 검색하는 기존 주제 검색의 성능을 크게 향상시키지 못하고 있다. 이에 본 연구는 블로그를 대상으로 의견 문서와 비의견 문서의 단어 통계를 비교 분석함으로써 의견 검색에 활용할 수 있는 통계적 특성을 파악하고자 한다. TREC의 블로그 트랙에서 사용했던 Blogs06 컬렉션과 150개의 TREC 토픽을 실험 데이터로 사용하였다. JS divergence를 이용하여 의견 문서에서의 단어 확률 분포 간의 상이성을 비교 분석하였으며, TREC 토픽의 유형 및 주제 영역별로 의견 문서를 구분하여 확률 분포의 차이점을 살펴보았고, 의견 단어별 확률을 비교 분석하였다. 실험을 통해 토픽별 특성을 고려한 의견 탐지 방법의 필요성, 토픽별 긍/부정 의견 단어 추출의 효과성, 유형과 주제 영역의 상호 보완적인 특징, 긍정 의견 단어 사용의 유의점 등을 알아내었다.

Abstract

Opinion retrieval which searches the opinions expressed in documents by users cannot outperform significantly yet traditional topical retrieval which searches the facts. Therefore, the focus of this paper is to identify the statistical characteristics which can be applied to opinion retrieval by comparing and analyzing the term statistics of opinion and non-opinion documents in the blog domain. The TREC Blogs06 collection and 150 TREC topics are used in the experiments. The difference between term probability distributions in opinion documents is measured by JS divergence, and the difference according to the topic types and topic domains is also investigated. Moreover, the term probabilities of opinion terms are analyzed comparatively. The main findings of this study include the following: it is necessary to consider the topic-specific characteristics for the opinion detection; it is effective to extract positive and negative opinion terms according to the

• 제1저자 : 한경수
• 투고일 : 2010. 07. 19, 심사일 : 2010. 08. 12, 게재확정일 : 2010. 09. 09.
* 성결대학교 컴퓨터공학부 전임강사

topics; the topic types are complementary to the topic domains; and special attention has to be given to the usage of the positive opinion terms.

▸ Keyword : 의견 검색(Opinion Retrieval), 의견 탐지(Opinion Detection), 의견 단어(Opinion Terms)

Ⅰ. 서 론

최근 인터넷 사용자들은 블로그(blog), 게시판, 뉴스 댓글 등을 통해 다양한 방식으로 의견을 표출하고 있으며, 이 의견 정보들은 다른 인터넷 사용자나 기업에게 유용한 정보가 되고 있다. 따라서 단순 사실에 대한 검색을 넘어 의견을 찾는 의견 검색(opinion retrieval)의 필요성이 대두되었으며, 그 중요성은 날로 커지고 있다.

의견 검색에 대해 정보 검색, 텍스트 마이닝, 자연어처리 분야에서 의견 검색, 의견 마이닝(opinion mining), 감성 분석(sentiment analysis) 등의 이름으로 다양한 연구가 진행 중이다. 특히, 미국 NIST(National Institute of Standards and Technology)에서 주관하는 TREC(Text REtrieval Conference)에서 2006년도부터 블로그 트랙(blog track)을 시작하면서 의견 검색에 대한 연구가 활발히 진행되고 있다.

그러나 블로그 트랙에 제출된 많은 의견 검색 시스템들의 검색 결과가 일반 사실 기반 검색, 즉 주제 검색(topical retrieval) 성능을 향상시키지 못하거나 향상시키더라도 향상 폭이 매우 미미한 것으로 나타났다[1,2].

의견 검색 성능은 왜 이렇게 좋지 못한 것일까? 본 연구의 동기가 된 물음이다. 본 연구는 의견 검색의 성능이 이처럼 낮은 원인을 파악하기 위해 블로그 문서들을 대상으로 의견 문서들에서의 단어 통계를 분석해봄으로써 의견 검색 문제 자체의 특성을 분석해보고자 한다. 이를 위해 TREC의 블로그 트랙에서 사용하였던 실험 데이터를 이용하여 의견 문서들에서의 단어 확률 분포를 비교 분석하여 몇 가지 흥미로운 사실들을 발견하였다.

본 논문의 구성은 다음과 같다. 2장에서는 일반적인 의견 검색 시스템에 대한 관련 연구를 살펴본다. 3장에서는 의견 문서의 단어 통계 분석 방법에 대해 설명하고, 4장에서 실험 결과와 이에 대한 분석을 기술한다. 마지막 5장에서 본 연구의 결론과 향후 연구에 대해 논한다.

Ⅱ. 관련 연구

본 논문에서 다루는 의견 검색은 사용자 질의에 연관되면서도 의견을 포함하고 있는 문서를 검색하는 것을 목적으로 한다. 이런 측면에서 기존의 일반적인 주제 검색의 틀에서 대상 문서 도메인만 블로그로 한정되어 있는 블로그 검색[3]과는 차이가 있다.

의견 검색에 대해 가장 활발히 연구가 진행되고 있는 TREC의 블로그 트랙을 위주로 관련 연구를 분석한다.

TREC 참여 시스템 대부분은 그림 1과 같이 2단계로 구성되어 있다. 기존의 일반 검색 시스템과 같이 주어진 주제(topic)에 적합한 문서를 검색하는 주제 검색 단계를 거친 후, 의견 탐지를 위한 여러 자질(feature)을 기반으로 문서 순위를 재조정하는 의견 탐지(opinion detection) 단계를 거쳐 의견 검색을 수행한다.

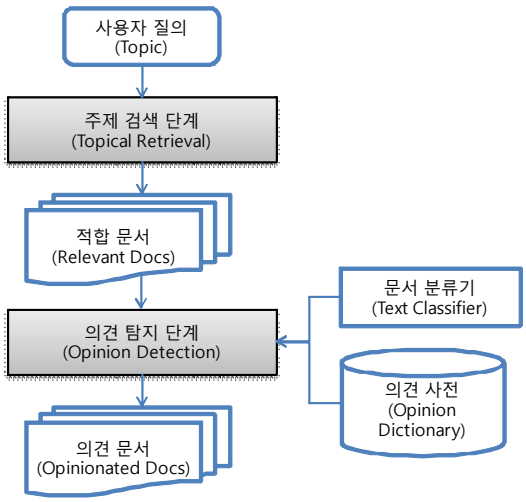

그림 1. 2단계 의견 검색 시스템 구조
Fig. 1. 2-Phase Structure for Opinion Retrieval

의견 검색의 대부분 연구에서 의견을 표현하는 단어 및 구절들을 모아 놓은 의견 사전(opinion dictionary)을 활용하는데, 대표적인 연구로 [4]을 들 수 있다. 이 연구에서는 의

견 탐지를 위해 의견 어휘 및 의견 연어(collocation) 등을 활용한다. 의견 어휘는 의견을 표현하는 블로그에서 상대적으로 자주 등장하는 단어로서 학습 데이터로부터 정보 이득(information gain)을 이용하여 추출되며, 신조어도 포함시키기 위해 학습 데이터의 저빈도 단어 중 일반 사전에 등재되지 않은 단어도 추출한다. 어휘 연어에는 IU(I/You) 연어와 AV(형용사-동사) 연어가 사용된다. IU 연어는 의견 표현 블로그에서 'I'나 'you'같은 대명사가 자주 사용된다는 사실에 기반하여, 학습 데이터로부터 'I believe', 'my assessment', 'good for you' 등의 연어를 추출하여 사용한다. AV 연어는 주관적인 형용사(예: 'good', 'bad', 'ugly')와 동사(예: 'hate', 'love', 'disgust')가 집중적으로 등장하는 문서는 의견을 표현하고 있을 것이라는 가정에 기반한다. 이 연어는 sm 유사도(distribution similarity) 'I bel구축된다. 이와 유사하며,주관적인 형용사 '수동으로 구축belI 한 연구도 있었다[5]. 또한 수동으로 구축의 소규모 의견 단어 'WordNet 학습 의어(synonym)와 반의어(antonym) 'I bel확장하는 방식어와 AV(사전을 구축b거나[6], SentiWordNet에 기반하여,주관적인 형용사 '추출하여 활용한 연구도 있다[7]. 한편, 의견 문서 및 비의견 문서에서의 단어 출현 통계를 이용하여 의견 단어에 가중치를 부여하기도 하였다[8,9]. 한국어에 대한 연구에서는 문장 단위로 긍/부정 정보가 태깅된 리뷰 문장으로부터 의견 사전을 추출하는 연구가 주류를 이루었다[10,11].

의견 탐지를 위해 학습 데이터를 구축하여 문서 분류 문제로 해결하는 방법론들도 있었다. 특정 주제에 대한 리뷰 사이트에서 주관적인 문서를, 위키피디어(Wikipedia)나 신문기사로부터 객관적인 문서를 수집하여 학습 문서로 활용하고 SVM(Support Vector Machine)이나[12] 지수회귀모형(logistic regression model)을 분류기로 이용하였다[7].

의견 탐지 단계에서 발견된 의견이 해당 주제와 연관되지 않은 것일 수도 있으므로 주제와 연관된 의견을 담고 있는 문서를 찾아야 한다. 이를 위해 의견과 주제 사이의 관계를 고려해야 하는데, 기존 연구들은 주제 단어(질의어)와 의견 단어가 출현한 위치를 기반으로 근접하여 등장한 경우 해당 의견은 주제와 연관된 것으로 간주하였다[5,12,13,14].

의견 문서의 순위 부여를 위해 주제 검색 단계의 적합성 점수와 주제 탐지 단계의 의견 점수를 결합하여야 한다. 기존 연구들에서는 선형 결합과 로그 결합 등이 제안되었으며[8,12], 단순한 점수 결합의 문제점을 극복하고자 이론적인 기반 위에 모델 차원에서 통합하려는 시도가 있었다[15]. 긍/부정에 대한 의견 극성(polarity)의 엔트로피를 순위 부여에 적용한 연구도 있었다[14].

전술한 바와 같이 블로그 트랙에 제출된 의견 검색 시스템들의 의견 검색 성능이 주제 검색의 성능을 크게 향상시키지 못하고 있다[1,2]. 2008년도 TREC의 의견 검색 태스크(opinion-finding task)에 참여한 19개 그룹 중 주제 검색에 비해 의견 검색의 성능이 향상된 그룹은 10개 그룹 뿐이며, 이 중 5% 이상의 성능 향상이 보고된 그룹은 5개 그룹 뿐이다[1]. 2009년도 TREC의 의견 검색 태스크(opinionated facet blog distillation task)에서는 의견 검색 성능이 가장 좋은 4개 그룹의 성능을 주제 검색과 비교한 결과 성능이 향상된 그룹은 하나도 없었다.

본 연구는 TREC 블로그 트랙의 실험 데이터를 사용하여 블로그 문서의 단어 통계를 분석함으로써 의견 검색 문제에 대한 이해를 높이는데 기여하고, 보다 효과적인 의견 검색을 위해 어떤 점을 고려해야하는지 살펴보고자 한다.

III. 의견 문서의 단어 통계 분석

3.1. 통계 분석의 포커스

본 논문은 의견 검색에서 활용 가능한 통계적 특성을 파악하고자 의견 문서의 단어 통계를 분석한다. 통계 분석의 주된 포커스는 다음과 같은 연구 질문들을 검토하는 것이다.

- 의견 문서는 비의견 문서나 적합 문서 등과 단어 확률 분포 측면에서 서로 상이한가?
- 토픽의 유형이나 주제 영역에 따라 의견 문서의 통계적인 특성이 서로 상이한가?
- 긍/부정 의견 단어의 확률 분포가 의견 문서, 비의견 문서, 긍정 문서, 부정 문서 등에서 차이가 있는가?

4장의 실험은 위 질문들에 답할 수 있도록 설계되었다.

3.2. 단어 확률 계산

문서에 포함된 단어들을 Krovetz Stemmer[16]를 이용하여 스테밍한 후 각 스템의 빈도를 계산한다. 오해의 소지가 없는 한, 본 논문에서는 스템과 단어를 편의상 혼용한다. 단어의 빈도에 기반하여 단어 확률은 최대 우도 추정(maximum likelihood estimation)으로 계산된다. 즉 단어 x의 확률 $p(x)$는 대상 문서 집합에서의 단어 x의 빈도 $C(x)$를 이용하여 다음과 같이 계산된다.

$$p(x) = \frac{C(x)}{\sum C(x)}$$

3.3. 확률 분포의 상이성 측정

본 논문에서는 문서에 등장하는 단어 통계에 기반하여 의견 문서 집합이 다른 문서 집합과 얼마나 상이한지를 알아보고자 한다. 이를 위한 척도로서 다음 식과 같은 Kullback-Leibler divergence(KL divergence)[17]를 생각해볼 수 있다.

$$D(q||r) = \sum_{x} q(x) \log \frac{q(x)}{r(x)}$$

이 수식은 두 확률분포 q와 r의 차이를 교차 엔트로피(cross entropy)를 이용하여 계산하는 방법으로서, $0 \log \frac{0}{r} = 0$, $q \log \frac{q}{0} = \infty$로 정의된다. 또한 $r(x) = 0$인 경우 평탄화(smoothing)가 필요하다. KL divergence는 교환법칙이 성립하지 않아서($D(q||r) \neq D(r||q)$) 본 실험의 주요 척도로 사용하기에는 어려움이 있다. 따라서 본 실험에서는 KL divergence를 수정 보완한 Jensen-Shannon divergence(JS divergence)[18]를 사용한다. JS divergence는 다음과 같이 계산된다.

$$\begin{aligned} A(q,r) &= D(q||\frac{q+r}{2}) + D(r||\frac{q+r}{2}) \\ &= 2\log 2 \\ &\quad + \sum_{x \in both} \begin{pmatrix} q(x) \log \frac{q(x)}{q(x)+r(x)} \\ + r(x) \log \frac{r(x)}{q(x)+r(x)} \end{pmatrix} \end{aligned}$$

JS divergence는 비교 대상 확률 분포로부터 가상의 평균 분포를 구하고, 각 확률 분포와 이 평균 분포 사이의 차이를 합하여 계산한다. 교환법칙이 성립하며, 값의 범위가 0부터 $2\log 2$까지로 제한된다.

IV. 실험 및 평가

4.1. 실험 데이터

실험에는 Blogs06 컬렉션을 사용하였는데, 이 컬렉션은 2005년 12월 6일부터 2006년 2월 21일까지 11주 동안 수집된 블로그 데이터[19]로서 TREC 블로그 트랙의 실험에 사용되었다. Blogs06 컬렉션은 총 크기가 148GB인데, 블로그 자체를 의미하는 XML 피드(38.6GB), 개개 블로그 문서와 커멘트를 의미하는 고유링크(permalink) 문서(88.8GB), 블로그의 첫페이지에 해당하는 HTML 홈페이지(28.8GB) 등 3가지 요소로 구성된다. 의견 검색 태스크에서는 고유링크 문서가 검색 대상이 되는데, 총 3,215,171개의 문서가 포함되어 있다. 2009년도 블로그 트랙에서는 Blogs08 컬렉션[20]을 사용하는데, 이 데이터 크기가 2TB를 넘는 대용량으로 실험에 어려움이 있어 본 실험에서는 Blogs06 컬렉션을 사용한다.

블로그 트랙의 여러 태스크 중 본 실험에서는 의견 검색 태스크를 대상으로 한다. 의견 검색 태스크는 특정 주제에 대해 블로거들이 어떤 생각을 가지고 어떤 의견을 표출하는지를 알아내는 것을 목표로 하여, 특정 주제에 적합하면서도 의견을 담고 있는 문서를 검색하는 작업으로 정의된다.

검색 질의(query)는 2006년도부터 2008년도 까지 3년 동안 TREC에서 사용했던 총 150개(851번~950번, 1001번~1050번)의 토픽(topic)에 대해 실험하였다[21]. 각 토픽에 대해 일반 주제 검색의 정답 집합으로 간주되는 적합 문서 집합, 토픽에 적합하면서도 토픽에 대한 의견이 기술되어 있는 의견 문서 집합, 토픽에 대해 긍정적으로 의견이 표현되어 있는 긍정 의견 문서 집합, 토픽에 대해 부정적인 의견이 포함되어 있는 부정 의견 문서 집합 등의 정보가 존재한다.

4.2. 의견 문서의 상이성

의견 문서 집합을 정의하는데 있어, 실험 데이터에 존재하는 전체 150개 토픽의 각 의견 문서 집합을 통합하여 하나의 의견 문서 집합으로 간주할 수도 있겠고, 각 토픽에 대한 의견 문서 집합을 별도로 구분 지을 수도 있겠다. 본 논문에서는 전자의 의견 문서 집합을 '통합 의견 문서 집합', 후자를 '개별 의견 문서 집합'이라고 칭한다.

표 1은 의견 문서 집합이 다른 문서 집합과 단어 확률 분포 측면에서 얼마나 상이한지를 JS divergence로 측정한 실험 결과이다. 표에서 통합 열은 통합 의견 문서 집합에 대한 실험 결과이며, 개별 평균 열은 150개의 개별 의견 문서 집합에 대해 실험하여 평균한 값이다.

표 1. 의견 문서 집합 확률 분포 비교
Table 1. Comparison on Probability Distribution of Opinion Document Sets

비교 대상	JS Divergence	
	통합	개별 평균
적합문서 : 부적합문서	0.0759	0.2347
의견문서 : 비의견문서	0.0873	0.2485
의견문서 : 적합문서	0.0295	0.2165
긍정의견문서 : 부정의견문서	0.0374	0.3589

실험 결과를 보면, 일반 주제 검색에서 다루는 적합 문서와 부적합 문서 사이의 상이도보다 의견 문서와 비의견 문서 사이의 상이도가 더 크다는 사실을 알 수 있다. 따라서 의견 문서와 비의견 문서를 구분하는 일은 일반 주제 검색에서 적합 문서와 부적합 문서를 구분하는 것보다 다소 높은 성능이 가능하리라 기대된다. 하지만 전체적인 의견 검색의 성능이 높아지려면 의견 문서 사이의 랭킹이 중요할 것이다.

한편, 의견 문서와 적합 문서 사이의 상이도는 상대적으로 크지 않으며, 의견 문서를 통합하여 측정할 경우 적합 문서와 매우 유사해짐을 알 수 있다. 즉 어떤 토픽에 대해 의견을 표현하는 문서에는 주제어들이 자주 등장할 것이기 때문에 의견 문서와 적합 문서는 단어 출현 통계가 비슷해진다. 따라서 적합 문서를 검색하는 주제 검색 후 의견 탐지를 통해 의견 문서를 검색해내는 방식의 2단계 의견 검색 시스템에서, 의견 탐지 방법이 토픽별 특성을 통합하여 일반화되어서는 좋은 성능을 기대하기 어려울 것이라고 추정할 수 있다. 다시 말해서, 의견 검색 시스템의 성능 향상을 위해서는 각 토픽별 특성을 별도로 고려해주는 의견 탐지 방법이 필요하다는 것이다.

긍정 의견 문서와 부정 의견 문서 사이의 상이도는 개별 의견 문서를 통해 측정하였을 때 가장 크게 나타났다. 그러므로 의견 문서들로부터 긍정 의견 단어나 부정 의견 단어를 추출하기 위해서는 일반적인 긍/부정 의견 문서를 수집하여 추출하는 것보다는 토픽별 긍/부정 의견 문서를 별도 수집하여 추출하는 것이 더 효과적일 것이라고 추정할 수 있다.

4.3. 토픽의 유형 및 주제 영역별 분류

실험 대상인 TREC 토픽은 다양한 종류의 유형과 주제를 포함하고 있다. 표 2는 토픽의 유형에 따라 분류한 결과이며, 표 3은 토픽이 다루는 주제 영역에 따라 분류한 것이다.

토픽은 총 12개의 유형으로 분류되었으며, 조직체, 인명, 저작품 등 특정 유형의 토픽이 상당수를 차지하여 고르게 분포되어 있지 않았다. 반면, 토픽의 주제 영역은 총 6개로 분류되었으며, 스포츠를 제외하면 나머지 5개의 주제 영역은 비교적 고르게 분포되어 있었다.

표 2. TREC 토픽의 유형별 분류
Table 2. Types of TREC Topics

유형	개수	TREC 토픽 예
조직체	39	Qualcomm, Mayo Clinic
인명	32	Steve Jobs, George Clooney
저작품	20	March of the Penguins, Big Love
상품/브랜드	18	Blackberry, Zyrtec
정책/법규	15	flag burning, China one child law
사건/행사	12	Cheney hunting, Davos
이슈	4	global warming, Oscar fashion
상	3	Sag Awards, Grammys
장소	3	Varansi, Bolivia
물질	2	cholesterol, lactose gas
관계	1	Abramoff Bush
주장	1	intelligent design

표 3. TREC 토픽의 주제 영역별 분류
Table 3. Topic Domains of TREC Topics

주제 영역	개수	TREC 토픽 예
과학기술	36	Qualcomm, cholesterol
엔터테인먼트	34	Big Love, Grammys
사회/종교	29	China one child law, scientology
경제	23	Whole Foods, World Bank
정치	22	Abramoff Bush, Ariel Sharon
스포츠	6	Olympics, Lance Armstrong

4.4. 토픽 유형별 의견 문서 상이성

토픽의 유형에 따라 의견 문서의 단어 출현 통계가 어떻게 다른지를 알아보는 실험을 시행하였다. 표 4는 토픽의 유형에 따라 의견 문서와 비의견 문서의 확률 분포 차이 및 긍정 의견 문서와 부정 의견 문서의 확률 분포 차이를 JS divergence로 측정한 결과이며, 의견 문서와 비의견 문서 사이의 divergence

값이 큰 값 순으로 정리하였다. 이 값은 각 유형별 토픽의 개별 의견 문서 집합에 대해 실험하여 평균한 값이다.

표 4. 토픽 유형별 확률 분포 차이
Table 4. JS Divergence according to Types of TREC Topics

유형	의견 : 비의견	긍정의견 : 부정의견
상품/브랜드	0.4034	0.4144
물질	0.3241	0.7909
조직체	0.2861	0.3721
상	0.2716	0.3245
이슈	0.2481	0.6885
저작품	0.2293	0.2906
장소	0.2115	0.2769
인명	0.1995	0.2650
사건/행사	0.1721	0.5062
정책/법규	0.1710	0.3187
주장	0.0868	0.1213
관계	0.0806	0.1619
평균	0.2237	0.3776
표준편차	0.0927	0.1987

1실험 결과를 보면, 상품/브랜드, 물질, 조직체, 상(award) 등의 토픽 유형이 유형 구분 없이 개별 측정된 의견 문서와 비의견 문서 사이의 상이도 평균값인 0.2485(표 1)보다 더 크게 나타났다. 또한 상이도가 유형에 따라 그 값이 크게 차이난다는 것을 알 수 있었다. 한편, 긍정 의견 문서와 부정 의견 문서 사이의 상이도가 유형 구분 없이 계산한 평균값 0.3569(표 1)보다 더 큰 유형은 물질, 이슈, 사건/행사, 상품/브랜드, 조직체 등이었다. 즉 이들 유형에 대해 긍/부정의 의견 표현이 비교적 잘 구분된다고 할 수 있겠다.

4.5. 토픽 주제 영역별 의견 문서 상이성

본 절에서는 토픽의 주제 영역에 따라 의견 문서의 단어 출현 통계를 살펴보는 실험에 관해 다룬다. 표 5는 토픽의 주제 영역에 따라 의견 문서와 비의견 문서의 확률 분포 차이 및 긍정 의견 문서와 부정 의견 문서의 확률 분포 차이를 JS divergence로 측정한 결과이며, 의견 문서와 비의견 문서 사이의 divergence값이 큰 값 순으로 정리하였다. 유형별 실험에서처럼 이 값은 각 주제 영역별 토픽의 개별 의견 문서 집합에 대해 실험하여 평균한 값이다.

표 5. 토픽 주제 영역별 확률 분포 차이
Table 5. JS Divergence according to Topic Domains of TREC Topics

주제	의견 : 비의견	긍정의견 : 부정의견
과학기술	0.3040	0.4465
경제	0.2814	0.3790
엔터테인먼트	0.2598	0.3269
스포츠	0.2418	0.2417
사회/종교	0.2024	0.3414
정치	0.1686	0.2855
평균	0.2430	0.3368
표준편차	0.0503	0.0716

주제 구분 없이 시행한 실험의 평균 상이도에 비해, 의견 문서와 비의견 문서의 상이도에서는 과학기술, 경제, 엔터테인먼트 영역의 상이도가 더 컸으며, 긍정 의견 문서와 부정 의견 문서 사이의 상이도에서는 과학기술과 경제 영역의 상이도가 더 크게 나타났다. 따라서 다른 주제 영역보다 과학기술 및 경제 영역의 의견 문서를 식별하기 더 용이하며, 긍/부정 의견의 구분도 상대적으로 명확하다고 할 수 있겠다.

표 4와 표 5에서 보듯이 주제 영역에 따른 의견 문서와 비의견 문서 사이의 상이도나 긍정 의견 문서와 부정 의견 문서 사이의 상이도는 유형에 따른 상이도에 비해 상대적으로 편차가 크지 않으면서도 비교적 안정적인 상이도 값을 유지한다. 반면, 유형에 따른 상이도는 몇몇 유형에 대해서는 상이도가 매우 크므로, 특정 유형에 대해서는 주제 영역보다 유형에 따라 의견 문서를 식별하는 것이 더 용이할 수 있을 것이다. 따라서 토픽의 유형과 주제 영역은 토픽별 특성을 고려할 때 서로 보완적인 역할을 수행할 수 있으리라 생각한다.

4.6. 의견 단어별 확률 비교

본 절에서는 의견 표현 단어에 대해 의견 문서 집합, 비의견 문서 집합, 긍정 문서 집합, 부정 문서 집합 등에서 단어 출현 확률이 어떻게 차이 나는지를 비교 실험하였다.

그림 2는 단어 'think', 'hope', 'opinion', 'view', 'critic', 'issue' 등에 대해 통합 의견 문서 집합 및 비의견 문서 집합에서의 단어 출현 확률을 비교한 것이다. 더불어 표 4와 표 5에서 의견 문서 집합과 비의견 문서 집합 사이의 상이도가 가장 컸던 상품/브랜드 토픽 유형과 과학기술 주제 영역에 대한 의견 및 비의견 문서 집합에서의 단어 출현 확률도 같이 비교하였다.

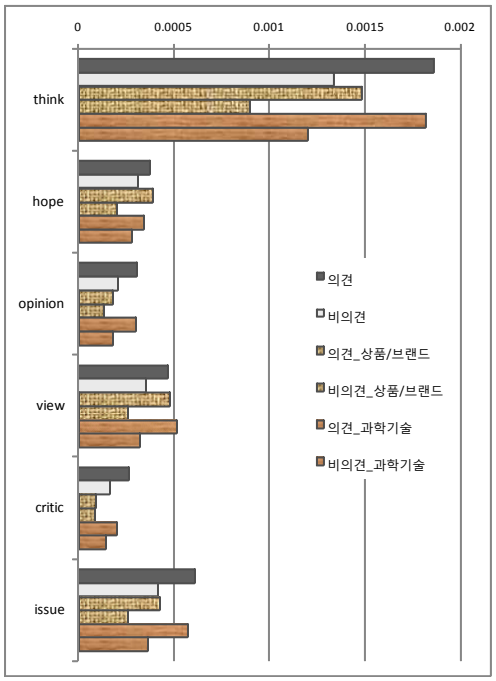

그림 2. 의견 단어별 확률 비교 (의견문서:비의견문서)
Fig. 2. Comparison of Opinion Terms Probability (Opinion Documents vs. Non-opinion Documents)

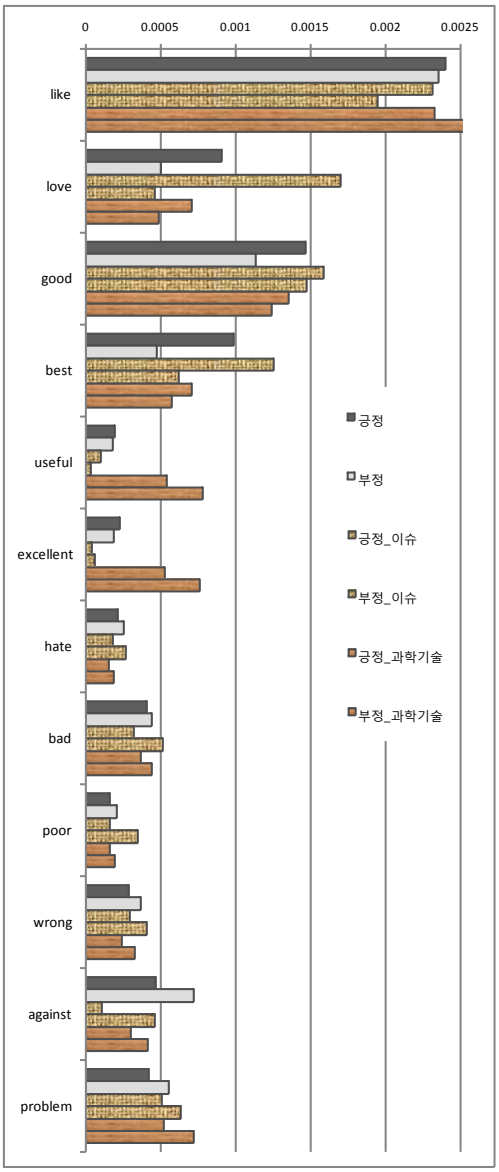

그림 3. 의견 단어별 확률 비교 (긍정의견문서:부정의견문서)
Fig. 3. Comparison of Opinion Terms Probability (Positive Opinion Documents vs. Negative Opinion Documents)

그림 2에서 보듯이, 일반적인 의견 단어들의 비의견 문서 집합에서의 단어 출현 확률 대비 의견 문서 집합에서의 단어 출현 확률의 상대적인 차이는 토픽 유형이나 주제 영역에 상관 없이 대부분 일관된 경향을 보인다.

논문의 가독성을 위해 토픽 유형과 주제 영역을 하나씩 선별하여 그림 2에 보였으나, 다른 토픽 유형과 주제 영역에 대해서도 그림 2와 유사한 실험 결과를 얻었다.

그림 3은 통합 긍정 의견 문서 집합과 부정 의견 문서 집합에 대해 단어 출현 확률을 비교한 결과이다. 실험에는 긍정 의견 단어로 간주되는 'like', 'love', 'good', 'best', 'useful', 'excellent' 등의 단어와 부정 의견 단어로 추정되는 'hate', 'bad', 'poor', 'wrong', 'against', 'problem' 등의 단어를 사용하였다. 이 실험 역시 표 4와 표 5에서 긍정 의견 문서 집합과 부정 의견 문서 집합과의 상이도가 가장 컸던 토픽 유형과 주제 영역에 대해 실험하였다. 물질 유형에 속한 토픽은 2개뿐이어서 다음으로 상이도가 높은 이슈 유형에 대해 실험하였다.

토픽 유형이나 주제 영역에 상관없이 통합한 긍정 의견 문서와 부정 의견 문서에 대해서는 예상했던 대로 긍정 의견 단어는 긍정 의견 문서에서 상대적으로 확률 값이 높고, 부정 의견 단어는 부정 의견 문서에서 상대적으로 확률 값이 높았다.

그러나 유형 및 주제 영역별 확률 값에는 일부 예외사항이 있었다. 과학기술 주제 영역에 있어서 'like', 'useful', 'excellent' 등의 확률 값이 긍정 의견 문서에서도 상당히 높으나, 긍정 의견 문서보다 부정 의견 문서에서 더 높게 나타났다. 이는 부정 의견을 표현하는 방식과 연관되는 것으로 추정 된다. 즉 부정적인 표현을 할 때, "I do not like~", "It is not useful/excellent~" 등으로 자주 표현하기 때문이다. 부정 의견 단어들은 예외 없이 부정 의견 문서에서 확률 값이 더 높게 나타나는 것도 이런 추정을 뒷받침한다. 따라서 긍/부정 의견 단어를 추출하거나 의견 탐지 단계에서 긍/부정 의견 단어를 사용할 때 긍정 의견 단어의 이런 특징을 고려해야할 것이다.

V. 결 론

본 연구에서는 TREC의 블로그 트랙에서 사용하는 Blogs06 컬렉션을 이용하여 블로그 문서의 단어 통계를 분석하였다. 이를 통해 다음과 같은 사실들을 발견함으로써 의견 검색 문제에 대한 이해를 높이는데 기여하였다. 첫째, 토픽별 의견 문서를 통합하면 적합 문서와 유사해지므로, 주제 검색 후 의견 탐지를 통해 의견 문서를 검색하는 2단계 의견 검색 시스템에서 성능 향상을 위해서는 각 토픽별 특성을 고려하는 의견 탐지 방법이 필요하다. 둘째, 토픽별 긍/부정 의견 문서를 개별적으로 측정하였을 때 상이도가 크므로, 토픽별 긍/부정 의견 문서를 별도로 수집하여 긍/부정 의견 단어를 추출하는 것이 효과적일 것이다. 셋째, 토픽의 유형 및 주제 영역별 의견 문서의 상이성을 계산해본 결과, 몇몇 유형에 대해서는 상이도가 매우 크고 주제 영역별 상이도는 편차가 크지 않으며 안정적인 상이도 값을 유지하므로, 유형과 주제 영역은 토픽별 특성을 고려하는데 상호 보완적인 역할을 수행할 수 있을 것이다. 넷째, 부정적인 의견을 표현할 때 긍정 의견 단어와 not 등의 부정어가 결합하여 표현될 수 있으므로, 긍/부정 의견 단어를 추출하거나 긍/부정 의견 단어를 사용하여 의견 탐지할 때 긍정 의견 단어의 쓰임에 주의해야 한다.

향후 연구로서 유형과 주제 영역 외에 토픽별 특성을 고려하는데 사용될 수 있는 기준이 있는지에 대한 추가 연구가 필요하며, 본 연구의 실험 결과 발견된 사항들을 고려하여 효과적인 의견 검색 모델을 설계하고자 한다.

참고문헌

[1] Iadh Ounis, Craig Macdonald, and Ian Soboroff, "Overview of the TREC-2008 Blog Track," Proceedings of the 17th Text Retrieval Conference (TREC-2008), Gaithersburg, Maryland, USA, Nov. 2008.

[2] Craig Macdonald, Iadh Ounis, and Ian Soboroff, "Overview of the TREC-2009 Blog Track," Proceedings of the 18th Text Retrieval Conference (TREC-2009), Gaithersburg, Maryland, USA, Nov. 2009.

[3] 신현일, 유은일, 류근호, "주제어 가중치 기법에 의한 효율적인 블로그 검색 시스템," 한국컴퓨터정보학회논문지, 제 15권, 제 4호, 1-9쪽, 2010년 4월.

[4] Kiduk Yang, Ning Yu, Alejandro Valerio, Hui Zhang, and Weimao Ke, "Fusion Approach to Finding Opinions in Blogosphere," Proceedings of the 1st International Conference on Weblogs and Social Media(ICWSM-2007), Boulder, Colorado, USA, Mar. 2007.

[5] Olga Vechtomova, "Using Subjective Adjectives in Opinion Retrieval from Blogs," Proceedings of the 16th Text Retrieval Conference (TREC-2007), Gaithersburg, Maryland, USA, Nov. 2007.

[6] Soo-Min Kim and Eduard Hovy, "Automatic Detection of Opinion Bearing Words and Sentences," Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005), pp. 61-66, Jeju Island, Korea, Oct. 2005.

[7] Ethan Zhang and Yi Zhang, "UCSC on TREC 2006 Blog Opinion Mining," Proceedings of the 15th Text Retrieval Conference (TREC-2006), Gaithersburg, Maryland, USA, Nov. 2006.

[8] Ben He, Craig Macdonald, Jiyin He, and Ladh Ounis, "An Effective Statistical Approach to Blog Post Opinion Retrieval," Proceeding of the 17th ACM Conference on Information and Knowledge Management (CIKM-2008), pp. 1063-1072, California, USA, Oct. 2008.

[9] 이승욱, 송영인, 임해창, "혼합 방식에 기반한 의견 문서 검색 시스템," 정보관리학회지, 제 25권, 제 4호, 115-129쪽, 2008년 12월.

[10] 남상협, 나승훈, 이예하, 이용훈, 김준기, 이종혁, "의견 어구 추출을 위한 생성 모델과 분류 모델을 결합한 부분 지도 학습 방법," 한국정보과학회 2008 종합학술대회 논문집, 제 35권, 제 1호(C), 268-273쪽, 2008년 6월.

[11] 주해종, 홍봉화, 정복철, "의견정보 모니터링을 위한 웹 마이닝 시스템에 관한 연구," 한국컴퓨터정보학회논문지, 제 15권, 제 1호, 149-157쪽, 2010년 1월.

[12] Lifeng Jia, Clement Yu, and Wei Zhang, "UIC at TREC 2008 Blog Track," Proceedings of the 17th Text Retrieval Conference (TREC-2008), Gaithersburg, Maryland, USA, Nov. 2008.

[13] GuangXu Zhou, Hemant Joshi, and Coskun Bayrak, "Topic Categorization for Relevancy and Opinion Detection," Proceedings of the 16th Text Retrieval Conference (TREC-2007), Gaithersburg, Maryland, USA, Nov. 2007.

[14] 윤홍준, 김한준, "오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법," 정보과학회논문지: 컴퓨팅의 실제 및 레터, 제 16권, 제 2호, 222-226쪽, 2010년 2월.

[15] Min Zhang and Xingyao Ye, "A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval," Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-2008), pp .411-418, Singapore, Jul. 2008.

[16] Robert Krovetz, "Viewing Morphology as an Inference Process," Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-1993), pp. 191-202, Pittsburgh, USA, Jun. 1993.

[17] Thomas M. Cover and Joy A. Thomas, *"Elements of Information Theory,"* Wiley-Interscience, New York, 1991.

[18] Lillian Jane Lee, *"Similarity-Based Approaches to Natural Language Processing,"* Phd Thesis, The Division of Engineering and Applied Sciences, Harvard University, May 1997.

[19] Craig Macdonald and Ladh Ounis, *"The TREC Blog06 Collection: Creating and Analysing a Blog Test Collection,"* DCS Technical Report TR-2006-224, University of Glasgow, 2006.

[20] The Blogs08 Test Collection, *http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html.*

[21] TREC 2008 Blog Track, *http://trec.nist.gov/data/blog08.html.*

저 자 소 개

한 경 수

1998 : 고려대학교 컴퓨터학과 이학사
2000 : 고려대학교 컴퓨터학과 이학석사
2006 : 고려대학교 컴퓨터학과 이학박사
2009-현재 : 성결대학교 컴퓨터공학부 전임강사
관심분야 : 정보 검색, 텍스트 마이닝, 자연어처리