지도적 잠재의미색인(LSI)기법을 이용한 의견 문서 자동 분류에 관한 실험적 연구

An Experimental Study on Opinion Classification Using Supervised Latent Semantic Indexing(LSI)

이지혜(Ji-Hye Lee)*
정영미(Young-Mee Chung)**

초 록

본 연구에서는 의견이나 감정을 담고 있는 의견 문서들의 자동 분류 성능을 향상시키기 위하여 개념색인의 하나인 잠재의미색인 기법을 사용한 분류 실험을 수행하였다. 실험을 위해 수집한 1,000개의 의견 문서는 500개씩의 긍정 문서와 부정 문서를 포함한다. 의견 문서 텍스트의 형태소 분석을 통해 명사 형태의 내용어 집합과 용언, 부사, 어기로 구성되는 의견어 집합을 생성하였다. 각기 다른 자질 집합들을 대상으로 의견 문서를 분류한 결과 용어색인에서는 의견어 집합, 잠재의미색인에서는 내용어와 의견어를 통합한 집합, 지도적 잠재의미색인에서는 내용어 집합이 가장 좋은 성능을 보였다. 전체적으로 의견 문서의 자동 분류에서 용어색인 보다는 잠재의미색인 기법의 분류 성능이 더 좋았으며, 특히 지도적 잠재의미색인 기법을 사용할 경우 최고의 분류 성능을 보였다.

ABSTRACT

The aim of this study is to apply latent semantic indexing(LSI) techniques for efficient automatic classification of opinionated documents. For the experiments, we collected 1,000 opinionated documents such as reviews and news, with 500 among them labelled as positive documents and the remaining 500 as negative. In this study, sets of content words and sentiment words were extracted using a POS tagger in order to identify the optimal feature set in opinion classification. Findings addressed that it was more effective to employ LSI techniques than using a term indexing method in sentiment classification. The best performance was achieved by a supervised LSI technique.

키워드: 의견 마이닝, 의견 자동분류, 개념색인, 용어색인, 잠재의미색인, 지도적 잠재의미색인
opinion mining, sentiment classification, LSI, supervised latent semantic indexing, latent semantic indexing, term indexing

* 연세대학교 문헌정보학과 대학원(cooljh99@naver.com) (제1저자)
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2009년 8월 18일 ■ 최초심사일자 : 2009년 8월 21일 ■ 게재확정일자 : 2009년 9월 3일
■ 정보관리학회지, 26(3): 451-462, 2009. 〔DOI:10.3743/KOSIM.2009.26.3.451〕

1. 서 론

'참여'와 '개방'을 키워드로 하는 웹 2.0 시대의 도래는 웹 환경을 바꿔놓았다. 과거 웹 1.0이 인터넷상의 정보를 모아서 제공하는 것에 그쳤다면, 웹 2.0은 인터넷을 이용하는 이용자라면 누구나 시간과 장소를 가리지 않고 데이터를 생성, 공유, 저장하고, 재생산할 수 있는 환경을 만들었다. 개인 미디어인 블로그 서비스를 통해 이용자의 콘텐츠 생산 활동이 활발해지고, 전자상거래의 활성화에 따른 웹상의 리뷰 문서 양의 증가로 자신의 의견이나 감정을 드러내는 의견 문서(opinionated document)가 폭발적으로 증가했다.

의견 문서의 증가와 함께 의견 정보가 가진 가치가 증가함에 따라 과거 텍스트의 자동 분류 영역에서 주제 분류가 주를 이루었던 것과는 달리 최근에는 텍스트가 담고 있는 의견에 주목하는 의견 마이닝(opinion mining) 연구가 활발히 진행되고 있다.

의견 마이닝은 데이터 마이닝 기법을 사용해 텍스트가 표현하고 있는 의견을 탐사하는 텍스트 마이닝의 하부 분야이다(Dave, Lawrence, and Pennock 2003). 기존의 텍스트 마이닝이 비구조화된 방대한 텍스트 집합으로부터 지식을 발견하는 과정이라면 의견 마이닝은 텍스트의 주제 정보가 아닌 텍스트가 가지고 있는 의견이나 감정 정보를 파악하는 과정이다. 자동으로 의견 정보를 추출하고 그것을 필요로 하는 대상에게 쉽게 활용할 수 있도록 제공하는 것이 의견 마이닝의 목적이라고 할 수 있다.

의견 마이닝은 새로운 지식 발견의 훌륭한 도구로 떠오르고 있으며 그 응용 분야는 매우 다양하다. 기업이나 정부에서는 자신들이 제공하는 상품이나 서비스에 대한 여론의 평가를 알아보는 행위가 필수적이다. 예컨대 의견 마이닝을 통해 과거 비싼 비용을 지불하며 행해왔던 평판 조사를 저렴한 비용으로 할 수 있다. 또 소비자 입장에서는 본인이 구입하고자 하는 물건을 먼저 구입한 사람들로부터 정보를 얻을 수 있고, 이것이 구매로 이어지기 때문에 기업에서는 의견 마이닝을 마케팅 전략을 세우는데 활용할 수 있다. 웹 광고의 경우 해당 제품에 대해 긍정적인 내용을 담고 있거나 경쟁사의 제품에 대해 부정적인 내용을 담고 있는 웹 페이지에 자사의 광고를 배치하는 전략을 택한다면 더욱 효과적일 것이다(Liu 2007).

현재까지 의견 마이닝에 관한 연구들은 문서의 객관성/주관성을 탐지하는 연구(Pang and Lee 2004; Yu and Hatzivassiloglou 2003), 문서가 긍정 혹은 부정 범주로 표현되는 문서의 의견 양극성(opinion polarity)을 판별하는 연구(Turney 2002), 강한 긍정이나 약한 부정과 같은 의견의 강도를 판단하는 연구(Wilson et al. 2004) 등 문서 단위의 연구를 포함하고 있다. 특히 의견 문서를 식별하고 대량의 데이터를 한꺼번에 확보하는 데 많은 시간과 비용이 들기 때문에 리뷰 문서를 대상으로 기계 학습 방법을 통한 의견 문서 자동 분류 연구가 활발하다(Dave, Lawrence and Pennock 2003; Chaovalit and Zhou 2005; Cui, Mittal and Datar 2006).

본 연구의 목적은 의견 문서 자동 분류에 개념색인의 하나인 잠재의미색인 기법을 도입하여 분류 성능을 향상시키는 데 있다. 이를 위해 문헌빈도 기준을 적용한 용어색인과 잠재의미

색인 및 지도적 잠재의미색인 기법을 각각 사용하여 분류 자질을 선정한 후 의견 문서 자동 분류 실험을 수행하였다. 형태소 분석을 위해 지능형 형태소 분석기를 사용하였고, 자동 분류에는 Weka 프로그램(http://www.cs.waikato.ac.nz/ml/weka/)이 제공하는 kNN 분류기를 사용하였다. 잠재의미색인 기법의 SVD 연산을 위해서는 MATLAB 7.0을 사용하였다.

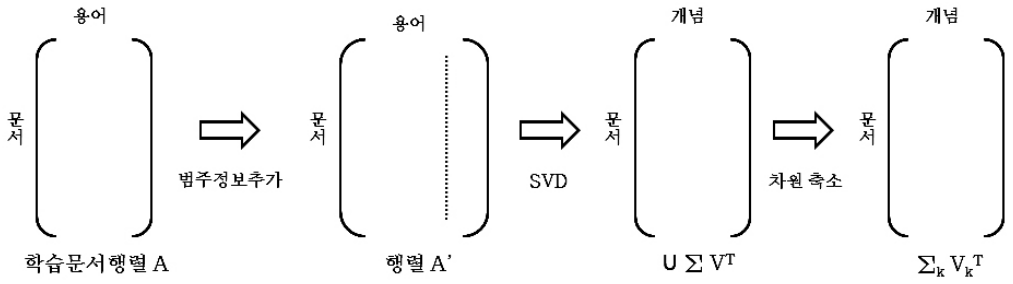
분류 실험은 먼저 지능형 형태소 분석기를 사용하여 품사에 근거한 분류 자질 집합을 생성하고, 용어색인, 잠재의미색인, 지도적 잠재의미색인 기법을 각각 적용하여 의견 문서를 자동 분류하는 과정으로 이루어진다.

2. 지도적 잠재의미색인

잠재의미색인 기법은 문서의 검색이나 필터링, 텍스트 범주화 등 다양한 분야에 적용되어 좋은 성능을 보여 왔다. 잠재의미색인 기법이 텍스트 범주화에 적용되는 방식은 전역적 잠재의미색인(global LSI)과 지역적 잠재의미색인(local LSI)의 두 가지가 있다. 전자는 곧바로 전체 행렬에 대해 SVD(singular value decomposition: 특이치분해) 연산을 적용하는 것으로, 학습 데이터에 범주 정보는 전혀 고려되지 않는 기법이다. 따라서 범주화 성능이 떨어지는 단점이 있다. 후자는 범주 정보를 효과적으로 적용해서 전역적 잠재의미색인보다 범주화 성능을 높이는 것이다.

잠재의미색인에서 n개의 문서 집합은 m x n 차원의 용어-문서 행렬 A로 표현된다. 행렬 A는 SVD 연산에 의해 3개의 특수한 행렬의 곱 $U\Sigma V^T$로 변환된다. SVD 연산 결과 생성된 행렬 U와 행렬 V는 각각 직교정규열을 갖는 m x n, r x n 크기의 행렬이고, 행렬 Σ는 r x r 크기의 대각선 행렬이다. SVD 연산에 의해 변화된 행렬 A는 축소된 k차원의 $U_k\Sigma_k V^T_k$로 변환되는데 이때의 k값은 보통 100에서 300 사이의 값을 선택한다(Dumais 1993; 정영미 2005; Ding 1999).

Chakraborti 등(2006)은 지도적 잠재의미색인인 '스프링클링(sprinkling)'기법을 제안했다. 기존의 잠재의미색인기법과 같은 과정을 따르되 문서 벡터에 SVD 연산을 적용할 때 문서의 범주 정보를 삽입하는 지도적 잠재의미색인 기법이다. 〈그림 1〉에서 보듯이 행렬을 SVD 하는 과정에서 학습문서에 범주를 의미하는 행렬



〈그림 1〉 스프링클링 과정

을 삽입함으로써 같은 범주에 속하는 것끼리 공통된 용어를 공유하는 효과를 갖는다. 이로 인해 서로 다른 범주에 속하는 문서 사이의 거리를 넓혀주는 효과를 볼 수 있다.

3. 의견 문서 자동 분류 실험

3.1 실험 개요

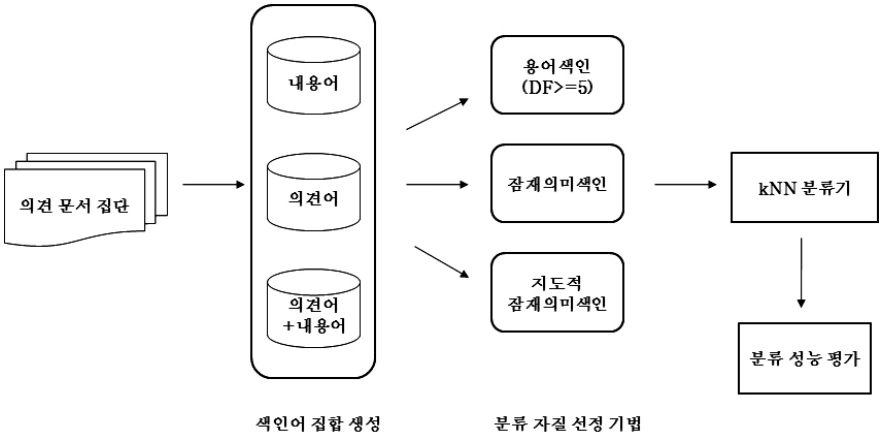
본 연구에서는 의견 문서 자동 분류에 내용어와 의견어 품사가 미치는 영향을 살펴보고, 지도적 잠재의미색인 기법을 도입하여 의견 문서 자동 분류의 성능을 향상시키고자 하였다. 실험 과정은 〈그림 2〉에 나와 있다.

(1) 색인어 추출 및 용어색인

웹에서 수집한 의견 문서 1,000건을 문서의 전처리 과정에서 지능형 형태소 분석기를 이용해 품사 태깅을 한 후 명사, 동사, 형용사, 부사, 어기를 추출한다. 내용어 집합은 명사, 의견어1 집합은 동사·형용사, 의견어2 집합은 동사·형용사·부사, 의견어3 집합은 동사·형용사·어기로 구성한다. 내용어와 의견어를 합친 자질 집합은 세 가지 의견어 자질 집합과 내용어 집합을 각각 통합한 것이다. 문서 자동 분류에서 문서를 표현하는 색인어는 분류 성능에 절대적인 영향을 미치므로 의견 자동 분류에서는 의견을 나타내는 품사의 추출이 필수적이다. 실제로 의견 자동 분류 시 의견을 나타내는 품사(형용사, 부사)를 추가하면 내용어 자질만 사용했을 때 보다 더 좋은 분류 성능을 낸다고 밝혀진 바 있다(황재원, 고영중 2008). 용어색인에서는 위에서 생성된 내용어, 의견어, 내용어+의견어 자질 집합에서 문헌빈도가 5 이상인 용어들만을 분류 자질로 사용한다. 용어 가중치로는 logTF와 logTF·IDF를 사용한 예비 실험결과 성능이 더 좋게 나온 logTF 가중치를 선택하였다.

(2) 잠재의미색인

내용어, 의견어, 내용어+의견어 등 세 가지



〈그림 2〉 의견 자동 분류 실험 과정

분류 자질 집합에 대해 잠재의미색인 기법을 적용하여 분류 자질을 선정한다. 세 가지 분류 자질 집합에 대한 SVD 연산에서 k 값으로 100, 150, 200, 300을 각각 사용하였다. 지도적 잠재의미색인은 용어-문서 행렬을 SVD 하기 전에 긍정과 부정의 두 범주 정보를 의미하는 2 x 1,000 행렬을 삽입한다.

(3) 의견 문서 자동 분류

용어색인, 잠재의미색인, 지도적 잠재의미색인 결과 각각 선정된 자질로 의견 문서를 표현한 후 *k*NN 분류기를 사용하여 분류한다. 일반적으로 범주화 실험에서 학습문서가 75%정도 되는 것을 참고하여 전체 문서 집단의 25%는 실험 집단으로 사용하고 나머지 75%를 학습 집단으로 사용하였다. 분류 성능 평가 척도로는 F_1 척도를 사용하였다.

3.2 실험 데이터

의견 문서 자동 분류를 위한 실험문서는 연구자가 직접 수집하였다. 수집 대상은 웹상의 의견 문서이며 제목을 제외한 본문을 수집하였다. 수집은 영화 리뷰 전문 사이트 무비스트(http://www.movist.com)와 리뷰 전문 서비스인 네이트 리뷰(http://review.nate.com), 네이버 뉴스 서비스(http://news.naver.com)에서 이루어졌다. 수집된 문서는 영화, 책, 음반, 공연, TV, 상품, 뉴스기사의 7가지 영역에서 긍정/부정 범주 각각 500건씩 총 1,000건이다. 리뷰 문서 수집 시 한 줄 평과 같이 한 문장 이하의 너무 짧은 글, 욕설이나 비속어, 이모티콘 같은 기호로 대부분 채워진 문서는 분류 실험 성능의 저하를 가져올 수 있으므로 수집 단계에서 제외하였다. 의견 문서의 긍정과 부정 범주의 판정은 연구자가 직접 문서를 읽고 판단하였다. 문서의 긍정과 부정 범주 판정에 있어 판정하기 애매하거나 어려운 것 역시 수집에서 제외하였다. 최종적으로 수집된 실험 문서 집단의 기본적 특성은 〈표 1〉과 같다.

전체 어절 수는 긍정문서 105,277개, 부정문서 54,413개로 긍정문서의 어절 수가 부정문서의 그것보다 두 배 가량 많았다. 어절 수나 문장 수로 볼 때 긍정문서의 문서 길이가 평균적으로 두 배 가량 길었다.

3.3 분류 자질 집합 생성

일반적으로 주제 자동 분류에서는 명사 형태의 내용어(content word)를 색인어로 추출하지만 본 실험에서는 의견을 표현하는 품사를 추출하였다. 예컨대 '좋은', '나쁜'과 같이 평가를

〈표 1〉 의견 문서 집단 특성

문서 수	긍정범주	부정범주
	500	500
전체 어절 수	105,277	54,413
문서 당 평균 어절 수	211	109
전체 문장 수	10,579	6,753
문서 당 평균 문장 수	22	14

포함하고 있는 형용사와 '짜증나다', '화나다'와 같이 감정 상태를 나타내는 동사를 의견어로 볼 수 있다. 반면 '부사'는 의견의 방향성(sentiment orientation)을 나타내기보다 '아주', '매우'와 같이 강도를 나타내는 것으로 판단되므로 의견어로서의 '부사'의 가치를 살펴 볼 필요가 있다. 또한 '어기(語基)'는 단어의 가장 중심이 되는 형태소로 어미와 직접 결합될 수 없고 자립형식도 아닌 것을 말한다. '지루하다'의 '지루', '심심하다'의 '심심', '뿌듯하다'의 '뿌듯'과 같이 감정 상태를 나타내는 어기를 의견어로 사용하는 것이 분류 성능에 도움이 되는지 알아보았다.

〈표 2〉에서 보듯이 추출된 전체 자질 수는 긍정문서가 88,749개, 부정문서가 46,610개로 긍정문서의 자질 수가 부정문서의 자질 수 보다 2배가량 많았다.

3.4 분류 자질 선정

분류 자질 선정은 문서 자동 분류의 성능에 큰 영향을 미친다. 원래 자질집합의 크기를 90%까지 축소해도 자동 분류 성능에는 큰 영향을 미치지 않는다는 것이 밝혀진 바 있다(Yang and Pedersen 1997). 따라서 성능 개선과 처리 속도 개선을 위해 자질 선정은 필수적이라 하겠다.

용어색인 기법을 사용한 실험에서는 문헌빈도가 5 이상인 용어만을 자질로 선정한 결과 〈표 3〉에서 보는 바와 같이 전체 키워드의 최대 39.1%부터 최소 22.6%로 자질이 축소되었다.

그러나 내용어와 의견어를 합친 자질 집합을 생성할 경우 내용어와 의견어 사이에 중복되는 키워드가 있었으므로 중복어를 제거하기 전과 제거한 후의 의견 자동 분류 성능을 살펴보았다. 문헌빈도 5 이상과 카이제곱 통계량 상위 20%를 기준으로 자질 선정을 한 후 중복어를 제거한 전과 후의 F_1 값을 각각 측정한 결과 중

〈표 2〉 분류 자질 집합 별 자질 수

범주	자질 집합	구분	자질 수
긍정 문서	내용어	일반 명사	55,549
		고유 명사	3,845
	의견어	동사	14,848
		형용사	5,921
		부사	6,717
		어기	1,869
	총 합		88,749
부정 문서	내용어	일반 명사	30,510
		고유 명사	2,118
	의견어	동사	6,910
		형용사	2,677
		부사	3,543
		어기	852
	총 합		46,610

〈표 3〉 문헌빈도를 이용한 자질 축소 결과(DF〉=5)

자질 집합		고유 키워드 수	선정된 자질 수	축소 비율
내용어	명사	12,136	2,748	22.6%
의견어	의견어1(동사·형용사)	1,756	686	39.1%
	의견어2(동사·형용사·부사)	2,615	987	37.7%
	의견어3(동사·형용사·어기)	2,214	812	36.7%
내용어+의견어	내용어+의견어1(명사·동사·형용사)	13,892	3,231	23.3%
	내용어+의견어2(명사·동사·형용사·부사)	14,751	3,391	23.0%
	내용어+의견어3(명사·동사·형용사·어기)	14,350	3,321	23.1%

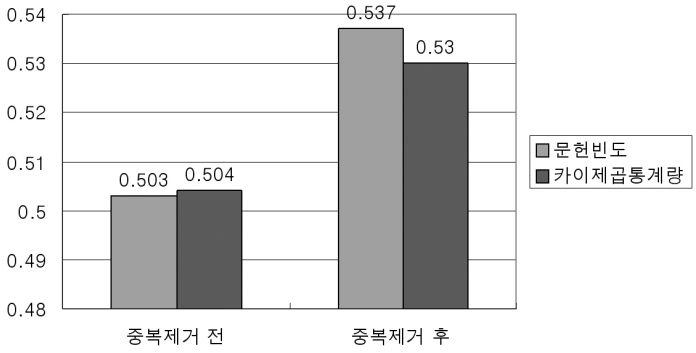
복어를 제거한 후의 분류 성능이 문헌빈도 사용 시 6.7%, 카이제곱 통계량 사용 시 5.1% 향상된 것으로 나타났다(그림 3 참조). 또한 중복되는 키워드가 전체 키워드 수에서 차지하는 비율이 〈표 4〉와 같이 1.3%~2.0%로 높지 않으므로 성능 향상과 효율성을 위해 중복되는 키워드는 내용어와 의견어를 합친 자질 집합에서 제거하였다.

잠재의미색인과 지도적 잠재의미색인 기법을 사용한 실험에서는 분류 성능을 높이고 처리 속도를 빠르게 하기 위해 장서빈도이고1인 자질을 제거하였는데 이 결과 〈표 5〉에서 보는바와 같이 전체의 60~71%로 자질이 축소되었다.

4. 실험 결과 분석 및 평가

4.1 색인 기법에 따른 분류 성능 비교

의견 문서 자동 분류를 위해 용어색인, 잠재의미색인, 지도적 잠재의미색인 기법으로 각각 자질을 선정한 결과 분류 성능을 비교하였다. 〈표 6〉은 문헌빈도 5 이상의 용어를 자질로 선정한 결과(DK-kNN)를 잠재의미색인(LSI-kNN) 및 지도적 잠재의미색인(SLSI-kNN) 기법의 분류 성능과 비교하여 성능 향상 정도를 나타낸 것이다. 지도적 잠재의미색인 기법으로 자질을 선정한 결과 평균 F_1 값이 0.898로 가장 높았다.



〈그림 3〉 중복 키워드 제거 전과 후의 F_1값 비교

〈표 4〉 내용어-의견어 간 중복 키워드 수

자질 집합		키워드 수	중복 키워드 수	중복비율
내용어+의견어	내용어+의견어1(명사・동사・형용사)	13,892	182	1.3%
	내용어+의견어2(명사・동사・형용사・부사)	14,751	297	2.0%
	내용어+의견어3(명사・동사・형용사・어기)	14,350	232	1.6%

〈표 5〉 장서빈도 1인 자질 제거 결과

자질 집합		키워드 수	선정된 자질 수	축소 비율
내용어	명사	12,136	7,284	60.0%
의견어	의견어1(동사・형용사)	1,756	1,246	71.0%
	의견어2(동사・형용사・부사)	2,615	1,782	68.1%
	의견어3(동사・형용사・어기)	2,214	1,553	70.1%
내용어+의견어	내용어+의견어1(명사・동사・형용사)	13,528	8,230	60.8%
	내용어+의견어2(명사・동사・형용사・부사)	14,157	8,574	60.6%
	내용어+의견어3(명사・동사・형용사・어기)	13,886	8,469	61.0%

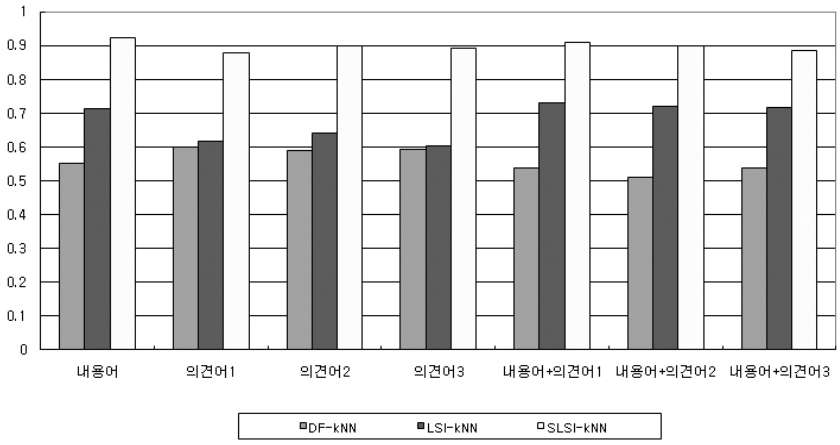
〈표 6〉 세 가지 색인 기법의 성능 비교

자질 집합	DF-*k*NN	LSI-*k*NN		SLSI-*k*NN		
	F_1값	F_1값	DF-*k*NN 대비 향상률	F_1값	DF-*k*NN 대비 향상률	LSI-*k*NN 대비 향상률
내용어	0.553	0.715	29.3%	0.924	67.1%	29.2%
의견어1	0.600	0.617	2.8%	0.879	46.5%	42.5%
의견어2	0.588	0.641	9.0%	0.900	53.1%	40.4%
의견어3	0.593	0.603	1.7%	0.891	50.3%	47.8%
내용어+의견어1	0.537	0.731	36.1%	0.908	69.1%	24.2%
내용어+의견어2	0.512	0.720	40.6%	0.900	75.8%	25.0%
내용어+의견어3	0.537	0.716	33.3%	0.884	64.6%	23.5%
평 균	0.560	0.677	21.8%	0.898	60.9%	33.2%

잠재의미색인 기법은 평균 F_1 값이 0.677로 뒤를 이었고, 문헌빈도 기준으로 자질을 선정한 결과는 0.560으로 가장 낮은 성능을 보였다. 지도적 잠재의미색인 기법을 사용했을 때, 평균 성능 향상률이 용어색인 대비 60.9%, 잠재의미색인 대비 33.2%로 분류 성능이 크게 향상되었음을 알 수 있다. 또한 지도적 잠재의미색인 기법을 사용했을 경우 모든 자질 집합의 분류 성능이 향상된 것으로 나타났다.

4.2 분류 자질 집합에 따른 분류 성능 비교

의견 문서 자동 분류 시 내용어, 의견어, 내용어+의견어 자질 집합 중 어떤 자질 집합이 가장 좋은 성능을 나타내는지 알아보기 위해 〈그림 4〉에서 내용어, 의견어, 내용어+의견어의 세 자질 집합 간 분류 성능을 비교하였다.



〈그림 4〉 자질 집합 간 분류 성능 비교

문헌빈도 기준으로 자질을 선정했을 때, 의견어1 집합의 F_1 값이 0.600으로 가장 성능이 좋았다. 잠재의미색인 기법을 사용했을 때는 내용어+의견어1 집합의 F_1 값이 0.731로 가장 높았으며, 지도적 잠재의미색인 기법에서는 내용어 집합의 F_1 값이 0.924로 가장 높게 나타났다. 자질 집합을 크게 내용어, 의견어, 내용어+의견어의 세 자질 집합으로 구분하여 각 색인 기법의 성능을 비교하면 문헌빈도 기준의 용어색인은 의견어 집합, 잠재의미색인은 내용어+의견어 집합, 지도적 잠재의미색인은 내용어 집합이 가장 좋은 성능을 보였다.

의견어 집합을 구성하는 용어의 품사로 동사·형용사 외에 부사와 어기를 추가하여 생성한 자질집합으로 분류한 실험 결과는 〈표 7〉과 같다.

의견어1 집합에 부사를 추가한 의견어2 집합은 문헌빈도 기준으로 자질을 선정했을 때는 의견어1 집합에 비해 성능이 하락했지만 잠재의미색인 기법 사용 시 의견어1 집합 대비 3.9%, 지도적 잠재의미색인 사용 시 2.4% 성능이 향상되었다. 반면 의견어1 집합에 어기를 추가한 의견어3 집합은 문헌빈도 기준, 잠재의미색인 기법 사용 시 모두 성능이 1.2%, 2.3%씩 하락하는 결과를 보였고 지도적 잠재의미색인을 사용했을 때만 1.4%의 성능 향상을 보였다.

내용어와 의견어를 결합하여 자질로 사용했을 때는 〈표 8〉에서 보듯이 부사, 어기를 추가

〈표 7〉 의견어 자질 집합의 성능 비교

	DF-*k*NN		LSI-*k*NN		SLSI-*k*NN	
	F_1값	향상률(%)	F_1값	향상률(%)	F_1값	향상률(%)
의견어1	0.600	-	0.617	-	0.879	-
의견어2(부사추가)	0.588	-2.0	0.641	3.9	0.900	2.4
의견어3(어기추가)	0.593	-1.2	0.603	-2.3	0.891	1.4

〈표 8〉 내용어+의견어 자질 집합의 성능 비교

	DF-*k*NN		LSI-*k*NN		SLSI-*k*NN	
	F_1값	향상률(%)	F_1값	향상률(%)	F_1값	향상률(%)
내용어+의견어1	0.537	-	0.731	-	0.908	-
내용어+의견어2(부사추가)	0.512	-4.6	0.720	-1.5	0.900	-0.9
내용어+의견어3(어기추가)	0.537	0	0.716	-2.1	0.884	-2.6

하는 것이 오히려 성능이 나빠지거나 그대로였기 때문에 성능 향상에 도움이 되지 않았다.

5 결 론

이 연구에서는 최근 크게 증가하고 있는 의견 문서를 분류하는 데 지도적 잠재의미색인 기법을 도입하여 의견 문서 자동 분류의 성능을 향상시키고자 하였다. 또한 의견 문서 자동 분류 시 분류 자질로 내용어(명사), 의견어(동사, 형용사, 부사, 어기), 내용어+의견어(명사, 동사, 형용사, 부사, 어기) 집합 중 어느 것이 가장 좋은 분류 성능을 내는지 알아보고 최적의 분류 자질 집합을 찾아내고자 하였다.

실험을 통해 밝혀진 의견 문서 자동 분류와 관련된 사실은 다음과 같다.

첫째, 의견 문서 자동 분류 시 지도적 잠재의미색인 기법으로 자질을 선정했을 때 용어색인의 문헌빈도 및 잠재의미색인 기법으로 자질을 선정했을 때와 비교하여 모든 자질 집합에서 분류 성능이 향상되었으며, 평균 F_1 값이 문헌빈도 대비 60.9%, 잠재의미색인 대비 33.2% 향상되었다. 잠재의미색인 기법을 사용했을 때도 문헌빈도 기반 용어색인에 비해 평균 21.8%의 성능 향상률을 보였다.

둘째, 의견 문서 자동 분류를 위한 최적의 분류 자질 집합은 문헌빈도를 기준으로 자질 선정을 했을 경우 의견어1 집합이 가장 좋은 성능을 보였다. 잠재의미색인 기법을 사용했을 때는 내용어+의견어1 집합이 가장 성능이 좋았으며, 지도적 잠재의미색인 기법은 내용어 집합에서 가장 좋은 성능을 보였다.

본 실험을 위해 수집된 의견 문서들은 7개의 영역에 걸친 다양한 주제 분야를 포괄하고 있다. 따라서 주제를 대표하는 내용어는 긍정, 부정 범주에 걸쳐 고루 분포되어 있어 용어색인 시 의견 범주를 식별하는 자질로는 성능이 떨어진 것으로 분석된다. 또한 내용어에 포함된 많은 동의어와 다의어 등이 걸러지지 않아 내용어 집합의 분류 성능을 저하시킨 것으로 보인다. 반면 잠재의미색인 기법 사용 시 내용어+의견어 집합이 가장 좋은 성능을 보였는데 이는 개념 색인 기법을 사용함으로써 주제어였던 내용어 간의 숨어있던 의미 구조를 더욱 정교하게 파악하게 되어 내용어와 의견어가 합쳐졌을 때 의견 문서 자동 분류에 적합한 자질 집합을 생성하였기 때문인 것으로 판단된다. 마지막으로 지도적 잠재의미색인 기법을 사용했을 때 내용어 자질 집합이 가장 좋은 성능을 보인 것은 실험 집단에 범주 정보를 삽입해줌으로써 잠재의미색인 시 부족했던 범주간의 의미구조

를 더욱 잘 표현하였기 때문일 것이다.

셋째, 의견어를 구성하는 품사를 동사·형용사 외에 '부사'와 '어기'를 추가한 실험에서 부사를 추가한 의견어2 집합은 개념색인인 잠재의미색인과 지도적 잠재의미색인 기법에서 분류 성능을 평균 3.15% 향상시켰으며, 어기를 추가한 의견어3 집합은 지도적 잠재의미색인 기법에서 1.4%의 성능 향상을 보였다.

결론적으로 의견 문서 자동 분류 시 개념 색인 기법을 사용하는 것은 분류 성능 향상에 효과적인 것으로 나타났다. 특히 범주 정보를 활용한 지도적 잠재의미색인 기법은 문헌빈도 기반 용어색인 대비 평균 60.9%의 높은 성능 향상률을 보여주었다. 그리고 의견 문서의 자동 분류에서 내용어, 의견어, 내용어+의견어 집합 등 자질 집합의 유형에 따라 분류 성능이 상이하게 나타나는 것으로 보아 분류 자질 집합의 구성이 분류 성능에 큰 영향을 미치는 것을 알 수 있다.

본 연구의 제한점으로는 특정 사이트에서 제공하는 7개의 영역에 한정된 의견 문서를 실험 문서로 사용하여 자동 분류를 수행하였으며, kNN 분류기만을 사용하여 실험한 결과를 바탕으로 성능을 비교하였다는 점을 들 수 있다. 따라서 장기간에 걸쳐 다양한 영역의 의견 문서들을 수집하여 실험문서 집단을 구성하고, 다양한 분류기를 사용하여 자동 분류 성능을 비교 연구할 필요가 있다. 또한 의견 문서를 작성한 이용자의 특성과 의견 문서에 나타난 언어 구조적인 측면을 고려한 포괄적인 의견 마이닝 연구가 필요할 것이다.

참 고 문 헌

정영미. 2005. 『정보검색연구』. 서울: 구미무역 출판부.

황재원, 고영중. 2008. 감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가. 『인지과학』, 19(4): 499-517.

Chakraborti, S., R. Lothian, N. Wiratunga, and S. Watt. 2006. "Sprinkling: supervised Latent Semantic Indexing." *Lecture Notes in Computer Science*, 3936: 510-514.

Chaovalit, P. and L. Zhou. 2005. "Movie Review Mining: a comparison between supervised and unsupervised classification approaches." *Proc. of the 38th Annual Hawaii International Conference on System Sciences, 2005.*

Cui, H., V. Mittal, and M. Datar. 2006. "Comparative experiments on sentiment classification for online product reviews." *Proc. of the 21st National Conference on Artificial Intelligenc,.* 1265-1270.

Dave, K., S. Lawrence, and D. M. Pennock. 2003. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." *Proc. of the 12th*

International Conference on World Wide Web, 519-528.

Ding, C. H. Q. 1999. "A similarity-based probability model for Latent Semantic Indexing." *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 59-65.

Dumais, S. T. 1993. "LSI meets TREC: A status report." *Proc. of the 1st Text REtrieval Conference(TREC-1)*, 137-152.

Liu, Bing. 2007. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer.

Pang, Bo., Lillian Lee, and Shivakumar Vaithyanathan. 2002. "Thumbs up? Sentiment classification using machine learning techniques." *Proc. of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 79-86.

Pang, Bo., and L. Lee. 2004. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." *Pro. of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain*, 271-278.

Turney, P. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, 417-424.

Wilson, T., J. Wiebe, and R. Hwa. 2004. "Just how mad are you? Finding strong and weak opinion clauses." *Proc. of the 2004 National Conference on Association for the Advancement of Artificial Intelligence*, 761-767.

Yang, Y. and J. O. Pedersen. 1997. "A comparative study on feature selection in text categorization." *Proc. of the 14th International Conference on Machine Learning*, 412-420.

Yu, H. and V. Hatzivassiloglou. 2003. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences." *Proc. of the 8th Conference on Empirical Methods in Natural Language Processing*, 129-136.