감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템

(A Korean Sentence and Document Sentiment Classification System Using Sentiment Features)

황 재 원 † 고 영 중 ††

(JawWon Hwang) (Youngjoong Ko)

요 약 최근 감정 분류에 대한 관심이 높아져 연구가 활발히 진행되고 있다. 문서 전체에 관한 감정의 분류도 중요하지만, 문서를 이루고 있는 문장에 관한 분류도 점차 그 필요성이 높아지고 있다. 본 논문에서는 한국어 감정 분류 시스템 구축을 위해서 추출된 한국어 감정 자질을 이용한 한국어 문장 및 문서 감정 분류에 관해 연구한다. 한국어 감정 분류의 시작은 감정을 내포한 대표적인 어휘로부터 시작하며, 이와 같은 감정 자질들은 문장 및 문서의 감정을 분류하는데 결정적인 관여를 한다. 한국어 감정 자질의 추출을 위하여 영어 단어 시소러스 정보를 이용하여 자질들을 확장하고, 영한사전을 통해 확장된 자질들을 번역함으로써 감정 자질들을 추출하였다. 추출된 감정 자질들을 사용하여, 단어 벡터로 표현된 입력문서를 이진 분류기인 지지벡터 기계(SVM: Support Vector Machine)를 이용하여 문장과 문서에 내포된 감정을 판단하고 평가하였다.

키워드 : 감정분류, 감정자질추출, 지지벡터기계

Abstract Sentiment classification is a recent subdiscipline of text classification, which is concerned not with the topic but with opinion. In this paper, we present a Korean sentence and document classification system using effective sentiment features. Korean sentiment classification starts from constructing effective sentiment feature sets for positive and negative. The synonym information of a English word thesaurus is used to extract effective sentiment features and then the extracted English sentiment features are translated in Korean features by English-Korean dictionary. A sentence or a document is represented by using the extracted sentiment features and is classified and evaluated by SVM (Support Vector Machine).

Key words : Sentiment Classification, Sentiment Feature Extraction, SVM

· 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-331-D00536)

· 이 논문은 제34회 추계학술대회에서 '감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 동아대학교 컴퓨터공학과
sftcap@gmail.com

†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr

논문접수 : 2007년 12월 7일
심사완료 : 2008년 3월 5일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제14권 제3호(2008.5)

1. 서 론

최근 초고속 인터넷이 폭 넓게 보급되고, 개인 컴퓨터 사용자의 연령층도 다양해져 많은 사용자들이 인터넷을 쉽게 이용할 수 있게 되었다. 최근 각광받는 컨텐츠 중의 하나가 UCC(User Created Contents)이다. UCC로 많은 멀티미디어 제작물들이 인터넷을 통해 생성되고 있으며, 그에 못지 않게 많은 텍스트 문서들도 쏟아져 나오고 있다. 이러한 텍스트로부터 추출할 수 있는 많은 정보 중에 유용한 정보의 하나가 작자가 해당 문서의 주제에 대해 표현한 감정 혹은 의견(sentiment or opinion)이다[1].

해당 텍스트 문서를 작성한 작자의 감정을 파악함으로써, 많은 응용 분야에 파악된 정보를 이용할 수 있다. 물건을 구입하려고 하는 소비자도 구입을 원하는 물건의 평판을 조사하여 구입여부를 판단하는데 도움을 얻을 수 있으며, 상품을 판매하는 기업은 자신의 상품의 평판을 조사하여 소비자의 불만사항이나 요구사항을 수렴하여 상품의 질을 개선시키고, 마케팅의 전략을 세우는데 조사된 정보를 이용할 수 있을 것이다. 일반적으로 인물이나 상품의 평판은 비싼 비용을 지불하고 조사(survey)되어 왔으나, 근래에 들어 인터넷을 통해 상품에 대한 평가(review)를 사용자들이 직접 입력하여 의견을 반영할 수 있는 시스템이 많이 보급되었다. 이렇게 자동 수집된 텍스트 문서들에서 자동으로 감정과 의견을 추출할 수 있다면, 저비용과 자동으로 원하는 정보를 얻을 수 있을 것이다. 최근 외국과 국내에서는 이러한 작자의 의견이 담겨있는 문서로부터 작자의 감정을 자동으로 판별하는 연구가 활발히 진행되고 있다.

지금까지의 문서 분류가 문서의 주제(topic)에 초점을 맞추었다면 감정 분류(sentiment classification)는 저자의 주제에 대한 긍정(positive)감정과 부정(negative)감정에 초점을 맞춘 연구 분야로서, 고객 평가의 요약, 공

공 의견 조사, 고객 성향 분석 등의 응용 영역을 가지고 있다.

문서 감정 분류에 효과적인 자질의 선정을 위해 고려해야 할 사항은 "감정 분류는 문서에 나타나는 단어의 형태가 아닌 단어의 의미에 기반 해야 한다"는 점이다. 감정 분류는 긍정과 부정의 감정에 초점을 두기 때문에 먼저 이를 가장 잘 표현하는 기본적인 자질의 생성이 중요하다. 본 연구에서는 영어 단어 시소러스의 유의어 정보를 이용하여 단어를 확장하고 이를 한영사전을 통해 번역하여 감정 자질을 생성하였다. 생성된 감정 자질을 이용하여 문서 감정 분류에 적용하고 감정은 문서 전체가 아닌 특정부분에 나타난다는 점을 고려하여 생성된 감정 자질이 자질의 수가 문서에 비해 적은 문장을 대상으로 평가하였다.

본 논문의 구성은 다음과 같다. 2장에서는 앞서 연구된 관련 연구에 대해 언급하였으며, 3장에서는 본 논문에서 문장 및 문서 감정 분류를 위한 자질 추출 방법과 한국어 감정 분류 시스템에 대해 언급한다. 4장에서는 한국어 감정 자질을 문장 및 문서에 적용하여 평가하였으며, 마지막 장에서는 결론 및 향후 과제에 대해서 기술한다.

2. 관련 연구

상품에 대한 평가와 영화에 대한 관객들의 평론에서 나타나는 주관적, 감정적 표현을 여러 기계 학습 방법과 자연어 처리 기술을 통해 문서를 분류하는 연구가 진행되고 있다.

특히 문서 감정 분류 시스템은 문서 분류의 특화된 분야이기 때문에 문서 분류에서 사용되어온 여러 가지 기계 학습 기법들이 문서 감정 분류에도 적용되어 왔다. 영화 평론과 상품 평가와 같은 특정 영역에서 나타나는 감정적 표현을 Naive Bayes, Maximum Entropy, Support Vector Machine등의 기계 학습을 통해 문서를 긍정과 부정의 범주로 분류하는 연구가 진행되어 왔다. 감정 분류에 대한 응용 영역으로는 먼저 상품에 대한 고객들의 평가에 들어있는 감정을 분류하여 내용을 요약하는 응용 분야(customer review)[2,3]와 공공의 의견을 조사하여 요약하는 응용 분야(public opinion survey)[4,5] 그리고, 고객들의 성향을 분석(trend analysis)[6]하는 분야 등 폭넓은 응용 영역을 가지고 있다.

또한, 분류의 대상이 문서뿐만 아니라, 문장[7,8], 구(phrase)[9,10], 토론의 연설가[11], 그리고 문장의 감정 패턴 분석을 통해 문장의 여러 감정적 표현을 인식하고 분류하는 연구도 수행되었다[12,13]. 그리고 국내에서는 최근 많은 문제가 되고 있는 인터넷상의 악성 댓글을 판별하는 시스템[14]에 관한 연구도 진행되고 있다.

3. 한국어 문장 및 문서 감정 분류 시스템

일반적인 문서 분류 시스템에서의 자질 선정 방법은 학습 문서에서 형태소 분석을 통해 내용어(content word)를 추출하고 추출된 대상 자질에 대해 가중치를 부여하는 것이 일반적이다. 하지만 감정 분류 시스템에서는 의미적 문서 분류를 위해서는 먼저 긍정과 부정을 나타내는 어휘인 감정 자질(sentiment feature)들을 따로 추출하여야 한다.

3.1 감정 자질 생성

감정 자질들과 일반적인 정보검색에서 사용되는 어휘들과의 가장 큰 차이점은 정보검색에서 사용되는 어휘들은 명사, 동사가 중요하게 사용되는 반면, 감정 분류의 감정 자질은 형용사, 부사들이 중요하기 때문에 그들 품사를 가지고 있는 단어들이 자질로 많이 포함된다는 점이다.

이러한 감정 어휘 집합을 추출하기 위해서는 여러 가지 어휘자원들이 필요한데 외국의 연구에서는 WordNet과 같은 어휘 의미망이 많이 사용되고 있다. 한국어의 감정 자질들의 확장을 위하여 한국어 사전을 파싱한 후 DB(Data Base)를 구축하여 동의어, 반의어 정보를 획득하고자 하였으나, 부정(13개), 긍정(12개)의 감정 자질만 획득하여, 원하는 결과를 얻을 수가 없었다. 한국어 사전에서는 어휘의 동의어와 반의어의 비중이 낮다고 판단하고 영어단어 시소러스 정보[15]를 이용하였다.

한국어 감정 자질 추출을 위하여 본 논문에서는 한국어에서 긍정과 부정을 나타내는 대표 어휘를 영어권 선행 연구 결과[12]를 바탕으로 대표 어휘를 선정하여 확장을 하였다. 본 논문에서 사용된 감정 자질의 내용은 아래와 같다.

- 감정 자질(Sentiment Feature)

부정 감정 자질(1,834개)이 긍정 감정 자질(861개)에 비해 약 2.1배 많이 생성되었기 때문에 영한번역 작업 시에 출현한 한국어 단어의 횟수를 사용하여 자질들의 수를 조정하였다. 2번 이상 출현한 자질들을 선정했을 때 부정 자질의 수가 844개로 긍정 자질(861개)의 수와 거의 균형을 이루었기 때문에 부정자질들은 2번 이상 출현한 자질들로 제한하였다.

최종적으로 생성된 감정 자질은 표 1과 같다.

표 1 자질 단어의 구성

자질 구분	내용
긍정 자질	좋다, 우수하다 등 861개의 자질 단어
부정 자질	나쁘다, 혐오하다 등 844개의 자질 단어

입력 문서를 형태소 분석 후, 앞 단계에서 선택된 자질을 기준으로 아래식의 TF-IDF 가중치 기법과 TF-ISF 가중치 기법을 적용한다.

TF-IDF 가중치 기법은 식 (2)와 같이 문서에 어휘 t가 나타난 어휘 빈도수(tf: term frequency) tf_t와 역 문서 빈도수(idf: inverse document frequency, 식 (1))의 곱으로 나타낸다.

$$idf_t = \log_2 \frac{N}{df_t} \tag{1}$$

$$weight_t = tf_t \cdot idf_t \tag{2}$$

여기서 N은 전체 문서의 수이며, df_t는 어휘 t가 출현한 문서의 수이다.

TF-ISF 가중치 기법은 문장에 출현한 어휘의 가중치를 계산하기 위해서 문서 빈도수 대신에 문장 빈도수를 사용하는 식을 말하며, 본 논문에서 문장 분류 시에 어휘 가중치 기법으로 사용된다. 식 (4)와 같이 문장에 어휘 t 가 나타난 어휘 빈도수(tf: term frequency) 와 역 문장 빈도수(isf: inverse sentence frequency), 식 (3))의 곱으로 나타낸다.

$$isf_t = \log_2 \frac{N}{sf_t} \tag{3}$$

$$weight_t = tf_t \cdot isf_t \tag{4}$$

여기서 N은 전체 문장의 수이며, sf_t는 어휘 t가 출현한 문장의 수이다.

문서 분류기는 지지 벡터 기계를 사용하였다. 지지 벡터 기계는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습 기법으로 두 개의 클래스의 구성 데이타들을 가장 잘 분리해 낼 수 있는 초평면(optimal hyperplane)을 찾는 모델이다. 지지 벡터 기계에서의 초평면은 식 (5)와 같이 나타낼 수 있다.

$$\vec{w} \cdot \vec{x} - b = 0 \tag{5}$$

여기서 $\vec{x}$는 분류하고자 하는 문서의 벡터이며 $\vec{w}$와 b는 학습 데이타로부터 학습되어 나온 결과이다. 학습 문서 집합을 $D=\{(y_i, \vec{x_i})\}$과 같이 나타냈을 때, 각각의 학습 문서 벡터$(\vec{x_i})$가 임의의 범주에 속한 문서이면 y_i의 값에 +1을 할당하고, 범주에 속하지 않은 문서에는 -1을 할당한다. 결국 지지 벡터 기계는 식 (6)과 식 (7)을 만족시키는 $\vec{w}$와 b를 찾는 문제이다.

$$\vec{w} \cdot \vec{x_i} - b \geq +1 \text{ for } y_i = +1 \tag{6}$$

$$\vec{w} \cdot \vec{x_i} - b \geq -1 \text{ for } y_i = -1 \tag{7}$$

지지 벡터 기계는 직선으로 나눌 수 있는 문제(linearly separable problem)에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 초평면을 설정하거나, 실제 데이타 벡터를 새로운 자질을 포함한 새로운 벡터 공간에 매핑하는 방법을 통해서 직선으로 나눌 수 없는 문제도 해결 할 수 있다. 본 논문에서는 Weka[16]에 제공된 SVM Toolkit을 사용 하였다.

3.3 실험 데이타

실험에 사용된 문서 데이타는 총 2,479개의 문서이며, 3개의 분야를 나누어 수집하여 아래의 표 2와 같이 신문기사 729개, 제품리뷰 395개, 영화리뷰 1,355개의 문서로 실험하였다. 실험에 사용된 문장 데이타는 총 53,762개의 문장이며, 3개의 분야를 나누어 수집하여 아래의 표 3과 같이 신문기사 13,733개, 제품리뷰 9,935개, 영화리뷰 30,094개의 문장으로 실험하였다. 모든 문장 및 문서를 사람이 직접 읽고 감정 여부를 판단하여 테스트 말뭉치를 구축하였다.

표 2 실험에 사용한 문서 테스트 말뭉치

분야	긍정	부정	총합
신문기사	417	312	729
제품리뷰	205	190	395
영화리뷰	703	652	1,355
총합	1,325	1,154	2,479

표 3 실험에 사용한 문장 테스트 말뭉치

분야	객관적 (Subject)	주관적 (Object)	총합
신문기사	5,926	7,807	13,733
제품리뷰	4,890	5,045	9,935
영화리뷰	10,831	19,263	30,094
총합	21,647	32,115	53,762

객관적(Object) 문장은 작자의 감정이 포함되지 않은 객관적인 문장이고, 주관적(Subject) 문장은 작자의 감정이 포함된 주관적인 문장을 지칭한다.

4. 실험 및 결과

4.1 성능평가 방법

본 논문에서는 다양한 자질 단어와 가중치 책정 방법을 사용하여 10-fold cross validation 방법으로 실험을 하였으며, 인터넷 사이트상에서 수집된 문서 집합의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확율(precision)과 재현율(recall)을 사용하였다.

정확율은 다음 식 (8)과 같이 표현된다.

$$\text{정확율} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{시스템이 적합하다고 판단한 문서수}} \tag{8}$$

재현율은 다음 식 (9)와 같이 표현된다.

$$\text{재현율} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{적합 문서수}} \tag{9}$$

정확율과 재현율을 하나의 값으로 표현해주기 위해서 다음 식 (10)과 같이 $F_1 - Measure$를 사용하였다.

$$F_1(r,p) = \frac{2 \cdot r \cdot p}{r+p} \tag{10}$$

식 (10)에서 r은 재현율에 해당하고 p는 정확율에 해당한다.

4.2 실험 결과

실험은 카테고리, 신문 기사, 제품 리뷰, 영화 리뷰의 3개의 카테고리로 나누어서 실험하였으며, 문장과 문서를 대상으로 실험을 하였다.

자질은 감정 자질을 사용하고 문서는 TF-IDF 가중치 기법을 사용했으며, 문장은 TF-ISF 가중치 기법을 적용하여 벡터로 표현하고 카테고리별로 분류를 수행하였다.

문장 및 문서 분류의 성능을 비교하기 위한 기준 시스템은 일반적인 정보검색에서 사용하는 내용어(명사, 동사, 형용사, 부사)를 사용하였고[2], 본 논문에서 제안한 시스템은 감정 자질을 사용하여 실험을 하였다.

4.2.1 문장 분류 실험 결과

실험의 성능은 아래의 표 4와 같다. 주관적 문장의 분류에서는 내용어에 비해 성능이 31.6% 향상되었고, 객관적 문장의 분류에서는 내용어에 비해 성능이 1.3% 하락하였다. 하지만 전체적으로 15.1%의 성능향상을 얻을 수 있었다.

표 4 문장 분류 성능

구분	감정	신문기사	영화리뷰	상품리뷰	평균
내용어	주관적	19.4	23.2	27.3	23.3
	객관적	70.6	74.1	67.7	70.8
감정 자질	주관적	47.0	51.4	66.2	**54.9**
	객관적	67.6	77.6	63.2	**69.5**

4.2.2 문서 분류 실험 결과

문서 분류는 긍정과 부정으로 분류하는 실험을 하였다. 문서 분류 실험의 성능은 아래의 표 5와 같다.

표 5 문서 분류 비교 실험 결과

구분	극성	신문기사	영화리뷰	상품리뷰	평균
내용어	부정	62.1	66.3	70.2	66.2
	긍정	71.7	68.4	71.4	70.5
감정 자질	부정	70.7	67.0	74.6	**70.8**
	긍정	77.7	68.7	77.1	**74.5**

표 5에서 보는 바와 같이 내용어에 비해 감정 자질을 사용한 결과, 부정에서는 4.6% 긍정에서는 4%의 성능향상을 얻었으며, 전체적으로 4.3%의 성능향상을 얻을 수 있었다. 표 4와 표 5의 성능은 F1-Measure의 값을 표시하였다.

이 같은 실험결과를 통해 감정 자질이 내용어에 비해 감정 분류에 더 유용하다는 것을 알 수 있다.

5. 결론 및 향후 과제

본 논문에서는 한국어의 감정자질을 이용해서 문장 및 문서의 감정을 분류하는 감정 분류 시스템을 제안하였다. 실험결과에서 보는 바와 같이, 제안된 시스템은 문장과 문서 분류에서 기준시스템보다 더 나은 성능을 보였다.

향후 연구로는 감정 자질과 그 주변의 단어와의 관계를 통해 자질을 확장하여 문장 분류에 적용할 것이고, TF-ISF 가중치 기법과 TF-IDF 가중치 기법과는 다른 방식으로 자질에 대한 가중치 부여 방법에 관해서도 연구를 수행할 것이다. 그리고 감정이 내포된 주관적 문장만을 대상으로 문서 분류를 하는 연구도 수행할 것이다.

참 고 문 헌

[1] M. Rimon, "Sentiment Classification: Linguistic and Non-Linguistic Issues," Hebrew University.

[2] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," EMNLP, pp. 79-86, 2002.

[3] K. Dave, S. Lawrence, D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," In Proceedings of WWW 2003, Budapest, Hungry, 2003.

[4] L.W. Ku, L.Y. Lee, T.H. Wu, and H.H. Chen, "Major Topic Detection and Its Application to Opinion Summarization," In Proceedings of the EMNLP conference, Geneva, 2004.

[5] S.M. Kim and E. Hovy, "Determining the Sentiment of Opinions," In Proceedings of the COLING conference, Geneva, 2004.

[6] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," In Proceedings of KDD'04, USA, 2004.

[7] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," In Proceedings of the ACL, pp. 271-278, 2004.

[8] Y. Mao and G. Lebanon, "Isotonic Conditional Random Fields and Local Sentiment Flow," In Proceedings of the NIPS, 2007.

[9] P. Turney, "Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews," In Proceedings of the ACL, pp. 417-424, 2002.

[10] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns," In Proceedings of the HLT/EMNLP, pp. 355-362, 2005.

[11] M. Thomas, B. Pang, and L. Lee, "Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts," In Proceedings of the EMNLP, pp. 327-335, 2006.

[12] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," ACM, pp. 617-624, 2005.
[13] E. Riloff and J. Wiebe, "Learning Extraction Patterns for Subjective Expressions," EMNLP, pp. 105-112, 2003.
[14] 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현", 한글 및 한국어 정보처리, pp. 285-289, 2006.
[15] http://eedic.naver.com/list_thesaurus.naver 네이버 영어단어 시소러스
[16] E. Frank, M. Hall, and L. Trigg, *Weka 3: Data Mining Software in Java*, The University of Waikato, 2006.