종자 어휘를 이용한 자질 추출과 지지 벡터 기계(SVM)을 이용한 문서 감정 분류 시스템의 개발

A Sentiment Classification System Using Feature Extraction from Seed Words and Support Vector Machine

황재원, Jaewon Hwang*, 전태균, Taegyun Jeon*, 고영중, Youngjoong Ko*
*동아대학교 컴퓨터공학과

요약 신문 기사 및 상품 평은 특정 주제나 상품을 대상으로 하여 글쓴이의 감정과 의견이 잘 나타나 있는 대표적인 문서이다. 최근 여론 조사 및 상품 의견 조사 등 다양한 측면에서 대용량의 문서의 의미적 분류 및 분석이 요구되고 있다. 본 논문에서는 문서에 나타난 내용을 기준으로 문서가 나타내고 있는 감정을 긍정과 부정의 두 가지 범주로 분류하는 시스템을 구현한다. 문서 분류의 시작은 감정을 지닌 대표적인 종자 어휘(seed word)로부터 시작하며, 자질의 선정은 한국어 특징상 감정 및 감각을 표현하는 명사, 형용사, 부사, 동사를 대상으로 한다. 가중치 부여 방법은 한글 유의어 사전을 통해 종자 어휘의 의미를 확장하여 각각의 가중치를 책정한다. 단어 벡터로 표현된 입력 문서를 이진 분류기인 지지벡터 기계를 이용하여 문서에 나타난 감정을 판단하는 시스템을 구현하고 그 성능을 평가한다.

핵심어: 감정분류(Sentiment Classification), 종자어휘(Seed Word), 지지벡터기계(SVM), 문서분류(Text Classification)

1. 서론

현재 발생되고 데이터베이스(Database)에 저장되는 데이터의 형태는 수치형 보다는 텍스트(text)가 훨씬 많다. 이러함 거대한 텍스트 집합으로부터 의미 있는 지식을 찾아내는 작업은 많은 분야에서 매우 다양하게 요구되고 있다. 텍스트로부터 추출할 수 있는 수많은 유용한 정보 중에 하나가 작자가 해당 문서의 주제에 대해 표현한 감정 혹은 의견(sentiment or opinion)이다[1]. 예를 들어, 상품개발자는 자신의 상품 혹은 경쟁사의 상품에 대한 평판을 아는 것이 상품개발과 마케팅을 위한 유용한 정보로 사용될 수 있으며, 또한 정부는 그들이 제시한 정책에 대한 시민들의 평가가 정책수행 및 새로운 정책수립을 위해서 유용하게 사용 될 수 있다. 전통적으로 이러한 상품이나 정책에 대한 평판은 비싼 비용을 지불하고 조사(survey) 되어 왔으나, 근래에 들어 인터넷을 통해 상품과 정책에 대한 평가(review)를 온라인으로 손쉽게 수집할 수 있게 됨에 따라, 텍스트로 저장된 평가 문서들에서 자동으로 감정과 의견을 추출할 수 있다면, 저비용으로 그리고 자동으로 의견조사가 가능할 것이다. 근래에 들어 외국에는 이러한 작자의 의견이 담겨있는 문서로부터 작자의 감정을 자동으로 판별하는 연구가 활발히 진행되고 있다.

전통적인 문서 분류(text classification)가 문서의 주제(topic)를 인식하고자 했다면 감정 분류(sentiment classification)는 저자의 주제에 대한 긍정감정과 부정감정을 인식하고자 하는 새로운 연구분야로서, 고객평가의 요약, 공공의견조사, 고객성향 분석 등의 폭넓은 잠재적인 응용영역을 가지고 있다.

일반적인 문서 분류는 사람이 문서에 나타난 자질을 보고 인식하여 정해진 범주로 분류하는 과정을 수학적으로 모델링하여 기계가 동일한 과정으로 학습하여 문서를 분류하도록 하는 것이다. 효과적인 문서 분류를 위해서 가장 중심이 되어야 하는 부분이 자질의 선정과 자질에 대한 가중치를 부여하는 방법이다.

문서 감정 분류는 문서에 나타나는 단어의 형태가 아닌 단어의 의미에 기반한다. 감정 분류의 초점이 되는 대상이 긍정과 부정이기 때문에 먼저 이를 가장 잘 표현하는 기본적인 단어인 종자 어휘를 생성한다. 종자 어휘로부터 문서 감정 분류를 위해서 사용될 충분한 양의 자질을 추출하기 위해 사전상의 유의어 및 반의어의 의미적 정보를 활용하여 단어의 의미를 확장시킨다. 확장된 자질을 이용하여 기존의 문서 분류 기법을 적용하여 문서에 대한 감정을 분류한다.

본 논문은 문서의 감정을 분류하기 위한 자질을 추출하기 위한 방법과 가중치 계산 방법에 대해 제안하고, 이를 통해 문서를 표현하여 기계학습 기법 중 하나인 지지 벡터 기계

(SVM)를 사용하여 문서의 감정을 분류하였다.

본 논문의 구성은 다음과 같다. 2장에서는 앞서 연구된 관련 연구에 대해 언급하였으며, 3장에서는 본 논문에서 제안하는 문서 감정 분류 시스템의 설계 과정에 대해 논의한다. 4장에서는 본 논문에서 제안하는 시스템의 성능을 평가하고 마지막 장에서는 결론 및 향후 과제에 대해서 기술한다.

2. 관련 연구

상품의 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 문서분류에서 사용되어온 여러 가지 기계학습 기법들이 문서 감정 분류에도 적용되어 왔다. 영화 평론과 상품 평가와 같은 특정 영역에서 나타나는 감정적 표현을 Naive Bayes, Maximum Entropy, Support Vector Machine 등의 기계 학습을 통해 문서를 긍정과 부정의 범주로 분류하는 연구가 진행되고 왔다[2,3,4,5].

감정 분류의 대한 응용 영역으로는 먼저 상품에 대한 고객들의 평가에 들어있는 감정을 분류하여 내용을 요약하는 응용분야(customer review)[2,6]와 공공의 의견을 조사하여 요약하는 응용분야(public opinion survey)[7,8] 그리고, 고객들의 성향을 분석(trend analysis)[9]하는 분야 등 폭넓은 응용 영역을 가지고 있다.

또한, 분류의 대상이 문서가 아니라 문장의 감정 패턴 분석을 통해 문장의 여러 감정적 표현을 인식하고 분류하는 연구도 수행되었다[10,11].

3. 종자 어휘를 사용한 문서 감정 분류 시스템의 설계

문서의 자질 선정 방법은 학습 문서에서 형태소 분석을 통해 내용어(content word)를 추출하고 추출된 대상 자질에 대해 가중치를 부여하는 것이 일반적이다. 하지만 의미적 문서 분류를 위해서는 먼저 긍정과 부정을 나타내는 어휘를 따로 추출하여야 한다. 이들 어휘들과 일반적인 정보검색에서 사용되는 어휘들과의 가장 큰 차이점은 정보검색에서 사용되는 어휘들의 품사는 명사, 동사가 중요하게 사용되는 반면 감정 분류에서는 형용사, 부사 등이 중요하게 사용된다는 점이다. 이러한 감정 어휘 집합을 추출하기 위해서는 여러가지 어휘자원들이 필요한데 외국의 연구에서는 WordNet[12]과 같은 어휘 의미망이 많이 사용되고 있다. 본 논문에서는 한국어에서 긍정과 부정을 나타내는 종자 어휘(seed words)를 생성하였고, 이를 대상으로 한글 유의어 사전을 통해 의미를 확장하여 긍정/부정 자질(feature)을 추출하고 가중치를 책정한다. 각 자질의 가중치는 문서 벡터의 생성에 적용되어 사용된다.

3.1 종자 어휘 생성 및 확장

문서의 긍정과 부정의 분류를 위해 각각의 의미를 나타내는 기본적인 종자 어휘를 생성한다. 종자 어휘의 생성은 영어권 선행 연구 결과[5,10]를 바탕으로 한국어 의미로 변환하여 2가지 종류의 종자 어휘를 생성한다. 생성된 종자 어휘는 한글 유의어 사전[13]을 통해 명사, 형용사, 부사, 동사를 대상으로 하며 목록에 더 이상 추가되는 단어가 없을 때까지 긍정과 부정에 대한 목록을 확장한다. 유의어는 동일한 의미 목록에 확장, 반의어는 반대 목록에 확장하는 것을 기본으로 한다. 이에 대한 내용은 [표 1]에 나타내었다.

[표 1] 종자 어휘와 자질 단어의 구성

구분	*내용*
종자어휘_1	{**긍정**:좋다 / **부정**:나쁘다}
종자어휘 _2	{**긍정**:좋다,우수하다,뛰어나다,괜찮다,명확하다, 행복하다,옳다 / **부정**:나쁘다,더럽다,낮다, 부정하다,그릇되다,가난하다,불운하다}
자질_1	종자어휘_1 의 유의어 단어 집합 {**긍정**:97 개 / **부정**:117 개}
자질_2	250 개의 학습데이타(신문기사)로부터의 감정 단어 추출
자질_3	자질_1 과 자질_2 의 단어 종합

확장된 단어를 자질로 선택하며 이를 대상으로 감정의 강약을 나타내기 위해 이미 구축된 감정 분류 데이터의 학습데이터를 이용해서 각 단어(t)마다 가중치(W_C)를 식(1)을 사용하여 추정한다. 각각의 긍정과 부정의 의미에 대하여 많이 등록이 된 단어가 의미가 더욱 강하므로 높은 가중치를 부여하며 전체에 비례하여 평준화한다. 아래는 긍정 일 때의 가중치 계산식이다.

$$W_{긍정} = \frac{P(긍정 \mid t)}{P(긍정 \mid t) + P(부정 \mid t)} \quad (1)$$

3.2 문서 표현 및 지지 벡터 기계(SVM)

입력 문서를 형태소 분석[14] 후, 앞 단계에서 선택된 자질을 기준으로 두 가지 방식의 가중치를 적용한다.

첫째, TFIDF 가중치 기법으로 계산된다.

둘째, 앞 단계의 TFIDF와 자질 단어의 가중치 식(1)의 곱으로 계산한다.

문서 분류기는 지지 벡터 기계(SVM)를 사용하였다.

지지 벡터 기계는 두개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습 기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 초평면(optimal hyperplane)을 찾는 모델이다[15]. 지지 벡터 기계에서의 초평면은 식(2)와 같이 나타낼 수 있다.

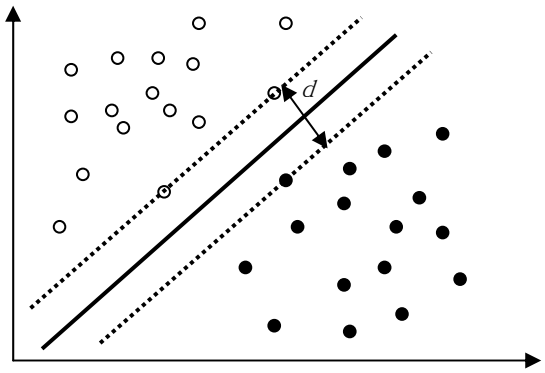
$$\vec{w}\cdot\vec{x}-b=0 \qquad (2)$$

여기서 $\vec{x}$는 분류하고자 하는 문서의 벡터이며 $\vec{w}$와 b는 학습 데이터로부터 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(3)과 식(4)를 만족시키는 $\vec{w}$와 b를 찾는 문제이다.

$$\vec{w}\cdot\vec{x}_i-b\geq+1 \; for \; y_i=+1 \qquad (3)$$

$$\vec{w}\cdot\vec{x}_i-b\geq-1 \; for \; y_i=-1 \qquad (4)$$

위의 수식들에 따르면 두 개의 클래스를 구분하는 초평면은 무수히 많이 존재하는데, 이들 초평면 들 중에서 최적의 초평면(Optimal hyperplane)은 두 클래스를 구분하는 거리(margin)가 최대가 되는 초평면으로 정의할 수 있다. [그림 1]은 두 클래스를 나누는 초평면 중에서 초평면들 사이의 거리(d)가 최대인 초평면을 보여주고 있다.



[그림 1] 초평면의 거리(Margin)

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 해결할 수 있다. 지지 벡터 기계 모델을 문서 범주화에 적용되어 좋은 성능을 보여왔다[16].

본 논문에서는 Weka[17]에 제공된 SVM toolkit을 사용하였다.

3.3 실험 데이터

문서 분류에 쓰인 데이터는 총 2517개의 문서이며, 3개의 분야를 나누어 수집하여 신문기사 731개, 제품리뷰 407개, 영화리뷰 1379개의 문서로 실험하였다. 사이트 상에 나타난 {찬성/반대}, {추천/비추천}의 정보를 바탕으로 사람이 직접 문서의 감정을 판단하여 테스트 말뭉치를 구축하였다.

[표 2] 실험에 사용한 테스트 말뭉치

분야	*긍정*	*부정*	*총합*
신문기사	418	313	731
제품리뷰	211	196	407
영화리뷰	671	708	1379
총합	1300	1217	2517

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 (5)과 같이 표현된다.

$$precision=\frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{시스템이 적합하다고 판단한 문서수}} \qquad (5)$$

재현율은 다음 식 (6)과 같이 표현된다.

$$recall=\frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{적합 문서수}} \qquad (6)$$

정확율과 재현율을 하나의 값으로 표현해주기 위해서 다음 식 (7)와 같이 F_1-measure를 사용하였다.

$$F_1(r,p) = \frac{2rp}{r+p} \qquad (7)$$

식 (7)에서 r은 재현율에 해당하고 p는 정확율에 해당한다.

4.2 실험 결과

실험 결과는 [표 3]에 요약되어 있다. 실험 결과를 살펴보면 먼저 자질은 종자어휘를 사용하여 확장될 때 성능이 향상되고 있으며 식(1)의 자질 가중치 방법을 TFIDF 기법에 적용했을 때가 더 많은 성능 향상을 얻을 수 있었다. 결과적으로 종자 어휘로부터 한국어 유의어 사전을 이용한 확장 자질 뿐만 아니라 학습 데이타로부터 추출한 모든 자질을 사용하고 식(1)의 자질 가중치 방법을 적용했을 때 가장 좋은 성능을 나타내었다.

[표 3] 자질과 가중치 추정 방법 선택에 따른 성능 비교

자질	가중치	긍정	부정	평균
자질_1	W(식 1)	0.61	0.61	0.61
자질_2	TFIDF	0.62	0.65	0.635
자질_2	TFIDF·W	0.70	0.59	0.645
자질_3	TFIDF	0.66	0.65	0.655
자질_3	TFIDF·W	***0.72***	***0.66***	***0.69***

5. 결론 및 향후 과제

본 논문에서는 신문 기사와 상품 및 영화 평가에 대해 감정을 분류하는 시스템을 제안하고 구현하였다. 문서 감정을 위해서 먼저 감정 분류를 위해 사용될 자질의 추출을 극히 제한적인 종자어휘를 사용해서 생성할 수 있는 방법을 제안했으며, 또한 추출된 자질들에 대해서 가중치를 추정하고, 이들 가중치를 이용해서 문서의 감정을 분류하였다. 제안된 시스템의 성능은 전체적으로 높은 성능이 나타나진 않았지만, 실험을 통해서 제안된 어휘 단어 목록의 확장 방법과 가중치 책정 방법에 따라 성능이 향상되는 것을 확인할 수 있었다.

향후 연구로는 한국어 유의어 사전뿐만이 아니라 전자 사전의 용례 등을 이용해서 감정 자질을 자동으로 추출하는 방식에 대한 연구를 진행 할 것이며, 감정의 특징상 문서 전체가 아닌 특정 문장에 강하게 표현되는 성질에 대한 연구를 수행할 것이다.

감사의 글

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

참고문헌

[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] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, “Sentimental Analyzer : Extracting Sentiments about a Given Topic using Natural Language Processing Techniques,” IEEE, p.427, November 19-22, 2003.

[4] N. Hiroshima, S. Yamada, O. Furuse and R. Kataoka “Searching for Sentences Expressing Opinions by using Declaratively Subjective Clues," ACL, pp.39-46, 2006.

[5] P. D. Turney and M. L. Littman, “Measuring praise and criticism: Inference of semantic orientation from association,” ACM, pp.315-346, 2003.

[6] 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.

[7] 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, 2002.

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

[9] M. Hu and B. Liu, “Mining and Summarizing Customer Reviews,” In Proceedings of KDD’04, USA, 2004.

[10] A. Esuli and F. Sebastiani, “Determining the Semantic Orientation of Terms through Gloss Classification,” ACM, pp.617-624, 2005.

[11] E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions,” EMNLP, pp.105-112, 2003.

[12] G.A. Miller, “Nouns in WordNet: A Lexical Inheritance System,” International Journal of Lexicography, Vol. 1, No. 4, pp.245-264, 1990.

[13] http://hanguli.cafe24.com, 우리말 음양낱말 이야기.

[14] 강승식, 한국어 형태소 분석 및 정보 검색, 홍릉과학출판사, 2002.

[15] V. Vapnik, The Nature of Statistical Learning Theory. Springer, New York, 1995.

[16] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In European Conference on Machine Learning(ECML),1998.

[17] E. Frank, M. Hall, and L. Trigg, Weka 3 : Data Mining Software in Java, The University of Waikato, 2006.