문장 감정 강도를 반영한 개선된 자질 가중치 기법 기반의 문서 감정 분류 시스템

(A Document Sentiment Classification System Based on the Feature Weighting Method Improved by Measuring Sentence Sentiment Intensity)

황 재 원 † 고 영 중 ††
(Jaewon Hwang) (Youngjoong Ko)

요 약 본 논문은 한국어 문서감정 분류에서 각 문장의 감정 정도의 차이를 고려하여 자질의 가중치를 계산하는 방법을 제안한다. 감정자질은 어휘 자원으로서 감정을 가지는 단어들의 집합이며, 학습데이터를 이용하여 이 감정자질의 카이제곱 통계량 값(χ^2 statistic)을 얻을 수 있다. 이렇게 얻어진 카이제곱 통계량 값으로 문서에서 출현한 각 문장의 감정강도를 수치화 할 수 있다. 각 문장의 감정강도는 문서에서 가장 강한 감정을 가진 문장에 대한 비율로 계산되며, 이 값을 TF-IDF 가중치 기법에 적용하여 최종적인 자질의 가중치를 결정하게 된다. 그리고 일반적으로 문서 분류에서 뛰어난 성능을 보여주는 지지벡터기계(Support Vector Machine)를 사용하여 기계학습을 수행한 후 성능을 평가한다. 성능평가에서 제안된 기법은 문장감정의 강도를 고려하지 않은 내용어(Content Word) 기반의 자질을 사용한 경우보다 약 2.0%의 성능향상을 얻었다.

키워드 : 감정 분류, 자질 가중치, 감정 자질, 감정 강도

Abstract This paper proposes a new feature weighting method for document sentiment classification. The proposed method considers the difference of sentiment intensities among sentences in a document. Sentiment features consist of sentiment vocabulary words and the sentiment intensity scores of them are estimated by the chi-square statistics. Sentiment intensity of each sentence can be measured by using the obtained chi-square statistics value of each sentiment feature. The calculated intensity values of each sentence are finally applied to the TF-IDF weighting method for whole features in the document. In this paper, we evaluate the proposed method using support vector machine. Our experimental results show that the proposed method performs about 2.0% better than the baseline which doesn't consider the sentiment intensity of a sentence.

Key words : Sentiment Classification, Feature Weighting, Sentiment Feature, Sentiment Intensity

1. 서 론

감정 분류는 텍스트(text)의 정보적인 부분이 아닌 감정적인 부분에 초점을 맞춘 연구 분야로서 최근 연구가 활발히 진행되고 있는 분야이다. 온라인(on-line)상에서 얻을 수 있는 텍스트의 양이 급증함에 따라 이 텍스트 집합으로부터 의미 있는 정보를 찾아내어 유용하게 활용하고자 하는 노력이 요구되고 있다. 이 중 활용할 수 있는 유용한 정보 중의 하나가 바로 해당 텍스트의 주제에 대한 의견이나 감정(opinion or sentiment)이다[1]. 예전에는 이러한 평판은 비싼 비용을 지불하고 조사(survey)되어 왔으나, 근래에 들어 인터넷을 통해 상품이나 정책 등에 대한 평가(review)를 온라인으로 손쉽

· 이 논문은 동아대학교 학술연구비 지원에 의하여 연구되었음

† 학생회원 : 동아대학교 컴퓨터공학과
sftcap@gmail.com
†† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr

논문접수 : 2008년 11월 13일
심사완료 : 2009년 4월 1일

정보과학회논문지 : 소프트웨어 및 응용 제36권 제6호(2009.6)

게 수집할 수 있게 됨에 따라, 텍스트 문서들로부터 자동으로 감정과 의견을 추출할 수 있다면, 저비용으로 그리고 자동으로 의견 조사가 가능할 것이다. 전통적인 문서 분류가 문서의 주제(topic)에 초점을 맞추었다면 감정 분류(sentiment classification)는 저자의 주제에 대한 긍정 또는 부정 감정에 초점을 맞춘 연구 분야로서, 고객 평가의 요약(customer review)[2,3], 공공 의견 조사(public opinion survey)[4,5], 고객 성향 분석(trend analysis)[6] 등의 응용 영역을 가지고 있다.

전통적인 자동 문서 범주화(automatic text categorization)는 미리 정의된 범주(category)에 문서를 자동으로 할당하는 기법과 관련된 연구 분야로서, 대량의 문서의 효율적인 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다.

자동 문서 범주화 과정은 문서를 어떤 자질을 통해 표현할 것인가를 다루는 자질 추출(feature extraction) 과정과 추출된 자질로 표현된 문서를 어느 범주로 할당할 것인가를 결정하는 문서 분류(text classification) 과정으로 구성된다. 감정 분류 역시 이러한 자동 문서 범주화 영역에 포함되는 영역이다.

자질 추출 과정에는 추출된 자질로 어떻게 문서를 표현할 것인가에 대한 색인(indexing)과정이 포함되며, 가장 일반적인 색인 방법은 벡터 공간 모델(vector space model)이다. 이 모델은 문장의 구분 없이 전체 문서에 출현한 각 자질의 빈도수(term frequency)를 가지고 표현하는 방법이다. 그러나 문서 내에 나타나는 문장들 중에는 해당 문서의 감정을 잘 나타내는 문장과 그렇지 못한 문장들이 있으며, 이러한 문장 감정 강도의 차이는 각 문장에 나타나는 감정 자질(sentiment feature)의 중요도에도 영향을 미친다. 그러므로, 본 논문에서는 자질 선택(feature selection) 기법 중 하나인 카이 제곱 통계량(χ^2 statistic)을 이용하여 감정 자질의 중요도를 얻고, 얻어진 카이 제곱 통계량 값을 이용하여 문장이 지닌 감정의 강도를 결정한다. 최종적으로 감정 자질이 어느 정도의 감정 강도를 지닌 문장으로부터 출현했는지를 색인 과정에 적용하고, 범주 별로 각 범주에 속하는 감정 자질을 강화하여 기계 학습과정에서 감정의 긍정과 부정에 대한 특징을 더 명확하게 학습하는 이점을 얻는다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 앞서 연구된 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 문장의 감정 강도를 고려한 감정 자질의 가중치 강화와 기계 학습 방법에 대해 자세히 설명한다. 그리고 4장에서는 실험 및 평가를 하며, 5장에서는 결론 및 향후 연구를 기술한다.

2. 관련 연구

문서 감정 분류는 문서 분류의 특화된 분야이기 때문에 문서 분류에서 사용되어 온 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 Machine 등의 기계 학습을 통해 문서를 긍정과 부정의 범주로 분류하는 연구가 진행되어 왔다[2,7-9]. 또한 분류의 대상이 문서뿐만 아니라, 문장[10,11], 구(phrase)[12,13], 토론의 연설가[14], 그리고 문장의 감정 패턴 분석을 통해 문장의 여러 감정적 표현을 인식하고 분류하는 연구도 수행되었다[15,16].

문서 분류에서 자질의 추출도 중요한 문제이다. 영어권 선행 연구에선 감정 분류에 적합한 자질을 추출하는 연구[5,15]가 수행되었고, 한국어 권 연구에서는 한국어 감정 자질을 추출하는 연구[17]가 수행되었다. 그리고 영어권 어휘 자원을 이용하여 감정 자질의 가중치를 결정하는 연구[18]도 진행되었다.

자질 추출 과정 중 자질 선택 단계는 문서에 나타난 여러 단어들 중 범주화에 유용하게 사용될 만한 단어들을 선택하는 과정으로 문서 빈도(document frequency), 상호 정보(mutual information), 카이 제곱 통계량(χ^2 statistic), 정보 획득량(information gain) 등의 기법이 있다.

자질 추출 과정 중 색인 단계는 선택된 자질을 통해 문서를 표현하는 단계로서, 일반적으로 벡터 공간 모델이 사용된다. 이 방법은 문서 전체에 나타난 자질들을 이용하여 문서를 하나의 벡터로 표현하는 방법으로 보통 자질의 빈도수와 역 문헌 빈도수(IDF)를 사용하여 문서를 표현한다. 그러나 이러한 기존의 방법은 문서가 가진 자질의 위치 정보나 문장 간 구분 등의 구조적 정보는 고려되지 못한다는 단점을 가진다.

이러한 한계를 극복하기 위하여 다양한 연구가 진행되었는데, 먼저 문서의 구조적 정보를 이용하기 위해 단어의 위치나 출현한 문장의 위치에 따라 가중치를 차등 적용한 방법이 연구되었으나, 모든 문서를 두괄식 또는 미괄식으로 가정하였기 때문에 신문 기사(article) 등 형식적인 문서를 제외하곤 그 적용이 힘들다[19]. 이런 약점을 보완하기 위해 제목과 문장 간의 유사도를 이용하여 중요한 문장을 결정하여 자질의 가중치에 적용하는 연구가 수행되었다[20].

3. 제안하는 접근법

3.1 전처리 과정

문장의 내용이나 특징을 잘 반영하는 단어를 내용어라고 하는데 본 연구에서는 내용어로서 형태소 분석의 결과 중 명사, 형용사, 동사, 부사만을 고려하였다. 전처

리 과정을 통해 입력 문서는 문장 단위로 내용어를 추출하게 되고, 추출된 내용어를 사용하여 문장 벡터들을 구성한다.

3.2 감정 자질

감정 자질은 감정을 가지는 단어의 어휘 자원으로 선행 연구[17]에서 추출한 감정 자질을 이용하였다. 하지만, 감정 자질의 질을 높이기 위하여 사람이 정제를 하여 감정이 약한 자질을 제거한 후 추가적인 확장을 하였다. 그 방법은 아래의 단계를 따른다.

1. 실험에 사용된 전체 문서의 형태소 분석결과 중 추출된 감정 자질 단어를 제외한 명사, 형용사, 부사, 동사의 단어를 추출
2. 한국어에 능숙한 2명의 주석자(annotator)가 각 단어의 긍정, 부정, 중립 여부를 태깅
3. 긍정 또는 부정이라고 태깅된 단어를 대상으로 우리말 국어 대사전 DB(data base)에 서 긍정의 동의어는 긍정으로, 반의어는 부정으로, 부정의 동의어는 부정으로, 반의어는 긍정으로 하여 단어를 추출

이렇게 확장된 감정 자질의 수는 아래의 표와 같다.

표 1에서 Senti는 선행 연구[17]에서 추출한 감정 자질을 정제한 감정 자질의 수이며, Senti DB는 이 추출된 감정 자질로 우리말 국어 대사전 DB에서 추출한 반의어, 동의어의 수이다. Annotation은 학습 문서에서 2명의 주석자가 태깅한 단어이며, Annotation DB는 주석자가 태깅한 단어를 질의(query)로 우리말 국어 대사전 DB에서 추출한 단어의 수이다.

이렇게 확장한 감정 자질은 긍정이 2175개, 부정이 3598개로 부정의 감정 자질이 더 많았기 때문에 부정 Annotation 단어를 대상으로 정규화를 수행하였다. 그 과정은 아래와 같다.

1. 학습 문서 내 DF가 1인 부정 Annotation 단어를 삭제(2155개에서 1042개로 감소)
2. 삭제되지 않은 부정 Annotation 단어로 우리말 국어 대사전 DB에서 반의어와 동의어 단어를 추출 (350개에서 223개로 감소)

N-Annotation은 정규화 된 부정 Annotation 단어이며, N-Annotation DB는 정규화 된 부정 Annotation DB 단어이다. 최종적으로, 긍정 감정 자질은 2175개, 부정 감정 자질은 2358개(Senti + Senti DB + N-Annotation + N-Annotation DB)를 추출하였다. 본 논문에서 사용된 감정 자질은 이 정규화 된 감정 자질이다.

표 2 정규화 된 부정 감정 자질의 수

구분	N-Annotation	N-Annotation DB	총계
부정	1042	223	1265

3.3 문장의 감정 강도 계산

문장의 감정 강도는 직관적인 방법에 의해 계산된다. 감정을 지닌 단어는 감정 자질을 통해 쉽게 알 수 있기 때문에 감정 자질이 많이 포함된 문장일수록 감정의 강도가 강하다고 생각할 수 있다.

문장 감정 강도는 식 (1)에 의해서 구한다.

$$Strength(S_i) = 1 + \frac{S_{i_{cnt}}}{D_{\max}}, \quad (1)$$

$Strength(S_i)$는 문장 감정 강도이며 D_{max}는 문서 D내에서 감정 자질을 가장 많이 가지는 문장의 감정 자질의 수이다. Si_{cnt}는 현재 문장의 감정 자질 수이다.

하지만 약한 감정을 가지는 단어를 많이 포함하고 있다고 해서 감정이 강한 문장이라고 보기 어렵기 때문에 감정 자질의 카이 제곱 통계량을 이용하여 문장 감정 강도를 구하는 방법도 사용하였다. 이 방법은 출현 횟수가 아닌 감정 자질의 카이 제곱 통계량의 수치를 합하여 문장의 감정 강도를 구하는 방법이다.

본 논문에서는 카이 제곱 통계량을 이용하여 문장 감정 강도를 구한다.

3.3.1 카이 제곱 통계량

감정 자질의 중요도를 카이 제곱 통계량으로 결정하였다. 카이 제곱 통계량을 구하기 위한 식은 다음과 같다.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}, \quad (2)$$

여기에서 A는 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있는 문서의 수, B는 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서의 수, C는 범주 c에 속해 있는 문서 중 용어 t를 포함하지 않은 문서의 수, 그리고 D는 범주 c에 속하지 않은 문서 중 용어 t를 포함하지 않은 문서의 수이다. 그리고 N은 전체 문서의 수이다.

각 범주 별로 얻어진 카이 제곱 통계량 값은 다음과 같은 식 (3)에 의해 가장 큰 값이 해당 용어의 자질 값

표 1 확장된 감정 자질의 수와 예

구분	Senti	Senti DB	Annotation	Annotation DB	총계
긍정	781	137	1123	134	2175
긍정 예	감격, 감동	순행, 낙관	강추, 찬사	청신호, 특효제	
부정	802	291	2155	350	3598
부정 예	불쾌, 혐오	흉조, 단점	미움, 실망감	질색, 암해	

이 되며 이 값을 감정 자질의 고유 가중치로 사용하여 식 (1)의 D_{max}와 Si_{cnt}를 감정 자질의 수가 아닌 감정 자질의 고유 가중치의 합으로 구하게 된다.

$$\chi^2\max(t) = \max_{i=1}^{m}\chi^2(t,c_i), \quad (3)$$

3.4 감정 자질 가중치 강화

문장의 감정 강도 계산에서 얻어진 문장의 감정 강도는 문서를 하나의 벡터로 표현할 때 감정 자질의 빈도수에 가중치를 강화하기 위해 사용된다. 문서에 출현한 자질의 빈도수는 각 문장에 출현한 자질의 빈도수의 합으로 구해진다. 이때 출현한 문장의 감정 강도에 따라 더해지는 빈도수의 수치가 달라지는데 이를 나타내는 식은 다음과 같다.

$$N(t|d) = \sum_{S_i \in d} tf(S_i,t) \times Strength(S_i), \quad (4)$$

위 식에서 $tf(Si,t)$는 문장 Si에서 출현한 감정 자질 t의 빈도수이며, $N(t|d)$은 문서 d에 출현한 문장 감정 강도에 의해 가중치가 강화된 감정 자질 t의 빈도수이다.

위 식에 따르면 각 감정 자질은 출현한 문장의 감정 강도(Strength)만큼의 가중치를 받게 되므로, 감정 강도가 강한 문장에서 나온 감정 자질은 실제로 출현한 빈도수보다 높은 값을 가지게 된다. 하지만 실제로는 모든 감정 자질의 가중치를 강화하는 것이 아닌, 각 범주에 해당하는 감정 자질만 강화하게 된다. 즉, 그림 1과 같이 부정 문서는 부정 감정 자질만을, 긍정 문서는 긍정 감정 자질만을 강화하게 된다.

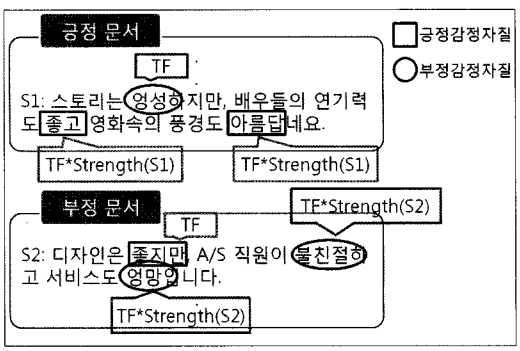

그림 1 범주 별 가중치 강화 예

그림 1의 예에서 보면 긍정 문서내의 긍정 감정 자질들은 문장 감정 강도 $Strength(S1)$만큼의 빈도수치가 증가하게 되고, 부정 감정 자질은 원래의 빈도수치만을 가지게 된다. 그리고 부정 문서내의 부정 감정 자질 또한 문장 감정 강도 $Strength(S2)$만큼의 빈도수치가 증가하게 되고, 긍정 감정 자질은 원래의 빈도수치만을 가진다. 이를 통해 한 문장에서 긍정 감정 자질과 부정 감정 자질이 동시에 출현했을 경우, 각 범주에 해당하는 감정 자질이 다른 범주의 감정 자질보다 기계학습 과정에서 더 기여하게 된다.

4. 실험 및 결과

4.1 실험 데이터

실험에 사용된 문서 데이터는 총 2,480개의 문서이며, 3개의 분야를 나누어 수집하여 신문기사 729개, 영화리뷰 1,356개, 상품리뷰 395개의 문서로 실험하였다. 모든 문서를 사람이 직접 읽고 감정 여부를 판단하여 테스트 말뭉치를 구축하였다.

표 3 실험에 사용한 테스트 말뭉치

분야	긍정	부정	총합
신문기사	417	312	729
영화리뷰	703	653	1356
상품리뷰	205	190	395
총합	1325	1155	2480

4.2 성능 평가 방법

본 논문에서는 5-fold cross validation 방법으로 실험을 하였으며, 인터넷 사이트상에서 수집된 문서 집합의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확률(precision)과 재현율(recall)을 사용하였다.

정확률은 다음 식 (5)와 같이 표현된다.

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

재현율은 다음 식 (6)과 같이 표현된다.

$$\text{재현율} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{적합 문서수}} \quad (6)$$

정확률과 재현율을 하나의 값으로 표현해주기 위해서 다음 식 (7)과 같이 $F_1 - Measure$를 사용하였다.

$$F_1(r,p) = \frac{2 \cdot r \cdot p}{r+p} \quad (7)$$

식 (7)에서 r은 재현율에 해당하고 p는 정확률에 해당한다. 본 논문에서는 F1-Measure값으로 실험결과를 표기한다.

4.3 분서 분류기

문서 분류기는 지지 벡터 기계를 사용하였다. 지지 벡터 기계는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습 기법[21]으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 초평면(optimal hyperplane)을 찾는 모델이다. 지지 벡터 기계에서의 초평면은 식 (8)과 같이 나타낼 수 있다.

$$\vec{w} \cdot \vec{x} - b = 0 \quad (8)$$

여기서 $\vec{x}$는 분류하고자 하는 문서의 벡터이며 $\vec{w}$와 b

는 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 (9)와 식 (10)을 만족시키는 $\vec{w}$와 b를 찾는 문제이다.

$$\vec{w}\cdot\vec{x_i}-b\geq+1 \text{ for } y_i=+1 \quad (9)$$

$$\vec{w}\cdot\vec{x_i}-b\leq-1 \text{ for } y_i=-1 \quad (10)$$

지지 벡터 기계는 직선으로 나눌 수 있는 문제(linearly separable problem)에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 초평면을 설정하거나, 실제 데이터 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 보여 왔다. 본 논문에서는 SVM Light[22]를 사용하였다.

4.4 실험 결과

실험은 실험 데이터 분야의 구분 없이 실험하였다. 먼저, 내용어를 사용한 실험에서는 75.31%의 결과를 얻었다. 본 논문에서는 이 결과를 기본 시스템으로 한다.

4.4.1 확장된 감정 자질 적합성 실험

확장된 감정 자질이 적합한 자질인지를 확인하기 위하여 감정 자질의 가중치 강화 없이 감정 자질이 포함된 문장만 대상으로 하여 감정 분류 실험을 하였다.

표 4 확장된 감정 자질 적합성 실험

대상 문장	모든 문장	감정자질포함문장
F1-Measure	75.31	75.36

감정 자질이 포함된 문장만을 대상으로 문서 분류 실험을 한 결과 미세하지만 나은 성능을 보였다.

4.4.2 문장 감정 강도 계산법 비교 실험

3.3절에서 설명한 문장 감정 강도 계산 방법 중 어느 방법이 더 적합한지를 알기 위하여 비교 실험을 수행하였다.

표 5 문장 감정 강도 계산법 비교 실험

구분	출현 횟수	Sum-chi
F1-Measure	75.68	76.62

출현 횟수는 단순히 감정 자질이 출현한 횟수를 더한 방법이고, Sum-chi는 카이 제곱 통계량 값을 이용하여 문서 감정 강도를 구한 방법이다. 실험 결과 단순히 감정 자질이 많이 나왔다고 해서 높은 중요도를 부여하는 방법보다는 감정의 강도가 강한 자질이 많이 나온 문장에 더 높은 중요도를 부여하는 방법이 더 낫다는 결과를 얻었다.

4.4.3 최종 실험 결과

제안한 방법의 실험은 아래의 단계로 각각 독립적으로 수행되었다.

방법1) 모든 문장을 대상으로 감정 자질 가중치 강화

방법2) 감정 자질이 포함된 문장만을 남겨 감정 자질 가중치 강화

방법3) 감정 자질이 포함된 문장만을 남겨 해당 범주에 속하는 감정 자질 강화

방법1의 실험결과는 표 6과 같다.

표 6 감정 자질 가중치 강화 실험 결과

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	76.62	+1.31

감정 자질의 가중치를 문장의 감정 강도를 고려하여 강화한 방법이 내용어를 사용한 기본 시스템보다 1.31% 나은 성능을 보였다.

방법2의 실험결과는 표 7과 같다.

표 7 감정 자질 포함 문장만 대상으로 실험

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	76.86	+1.55

감정 자질이 포함된 문장만 대상으로 감정 자질 가중치 강화를 수행한 결과 기본 시스템보다 1.55% 성능 향상을 보였다.

방법3의 실험결과는 표 8과 같다.

표 8 범주별 감정 자질 강화 실험

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	77.23	+1.92

각 범주에 해당하는 감정 자질만을 강화한 결과 기본 시스템보다 1.92% 향상된 성능을 보였다.

최종 성능 비교는 그림 2와 같다.

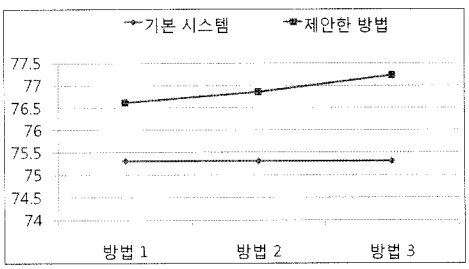

그림 2 최종 성능 비교

감정 자질이 포함된 문장만을 대상으로 문장 감정 강도를 고려하여 감정 자질의 가중치를 강화한 후, 각 범주에 해당하는 감정 자질만을 강화하여 학습한 방법3이 최대 1.92%의 성능 향상을 보였다.

5. 결론 및 향후 과제

본 논문에서는 문장의 감정 강도를 고려한 감정 분류 시스템을 제안하였다. 미리 구축된 어휘 자원인 감정 자질을 이용하여 문장의 감정 강도를 구하였다. 문장의 감정 강도를 효과적으로 구하기 위하여 학습 문서내의 감정 자질의 카이 제곱 통계량을 이용하였다. 이렇게 구해진 문서내의 문장 감정 강도의 값을 색인 과정에서 각 감정 자질의 빈도에 차등 적용하였다. 그리고 각 범주에 해당하는 감정 자질만을 강화하여 기계 학습 과정에서 각 범주의 특징을 명확하게 학습하는 이점을 얻었다.

제안한 방법을 사용했을 경우, 단순히 문서 전체에 출현한 단어의 빈도수를 이용하여 문서를 표현했을 때 보다 약 1.92%의 성능 향상을 얻을 수 있었다.

향후 과제로는 감정 표현의 이중 부정에 관한 패턴을 파악하여 파악된 패턴을 적용할 수 있는 방법에 관한 연구와, 문서가 아닌 문장의 감정 분류에 관한 연구도 수행할 것이다. 즉, 문서를 이루는 문장의 분류를 우선적으로 수행하여 해당 범주에 속하는 문장만을 대상으로 문서로 확장하는 방법에 관한 연구를 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," In *Proceedings of the 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 the 12th WWW*, pp. 519-528, 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 ACM SIGIR*, pp. 627-628, 2005.

[5] S.M. Kim and E. Hovy, "Determining the Sentiment of Opinions," In *Proceedings of the COLING conference*, pp. 1367-1373, 2004.

[6] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," In *Proceedings of the KDD*, pp. 168-177, 2004.

[7] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentimental Analyzer : Extracting Sentiments about a Given Topic using Natural Language Processing Techniques," In *Proceedings of International Conference on Data Mining*, pp. 427-434, 2003.

[8] N. Hiroshima, S. Yamada, O. Furuse and R. Kataoka, "Searching for Sentences Expressing Opinions by Using Declaratively Subjective Clues," In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 39-46, 2006.

[9] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," In *Proceedings of the ACM Transactions on Information Systems*, pp. 315-346, 2003.

[10] 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.

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

[12] P. Turney, "Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews," In *Proceedings of the ACL*, pp. 417-424, 2002.

[13] 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.

[14] 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.

[15] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," In *Proceedings of the CIKM*, pp. 617-624, 2005.

[16] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," In *Proceedings of the EMNLP*, pp. 105-112, 2003.

[17] 황재원, 고영중, "감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가", *한국정보과학회논문지: 컴퓨팅의 실제 및 레터*, 제14권 제3호, pp. 336-340, 2008.

[18] A. Esuli and F. Sebastiani, "PageRanking WordNet Synsets: An Application to Opinoin Mining," In *Proceedings of the ACL*, pp. 424-431, 2007.

[19] M. Murata, Q. Ma, K. Uchimoto, H. Ozaku, H. Isahara, and M. Utiyama, "Information Retrieval Using Location and Category Information," *Journal of the Association for Natural Language Processing*, Vol.7, No.2, 2000.

[20] Y. Ko, J. Park, and J, Seo, "Automatic Text Categorization using the Importance of Sentences," In *Proceedings of the 19th International Conference on COLING*, pp. 474-480, 2002.

[21] C. Cortes and V. Vapnik "Support-Vector Networks," *Machine Learning*, Vol.20, pp. 273-297, 1995.

[22] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many relevant Features," In *Proceedings of the ECML*, pp. 137-142, 1998.

황 재 원

2007년 동아대학교 컴퓨터공학과 학사
2009년 동아대학교 컴퓨터공학과 석사
관심분야는 자연어처리, 텍스트마이닝, 정보검색, 문서감정분류 등

고 영 중

1996년 서강대학교 수학과 학사. 1996년~1997년 LG-EDS 근무. 2000년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년~현재 동아대학교 컴퓨터공학과 조교수. 관심분야는 자연어처리, 텍스트마이닝, 의견마이닝, 정보검색, 대화시스템, 소프트웨어공학 등