http://dx.doi.org/10.7236/JIIBC.2013.13.5.27
JIIBC 2013-5-4

기계학습을 이용한 SNS 오피니언 문서의 자동추출기법

Automatic Retrieval of SNS Opinion Document Using Machine Learning Technique

장재영*

Jae-Young Chang

요 약 최근 들어 SNS가 대중화됨에 따라, 이들로 부터 오피니언을 분석하여 특정 이슈에 대한 여론을 파악하려는 다양한 연구가 진행되고 있다. SNS 환경에서 오피니언 분석을 위해서는 우선 게시글 중에서 오피니언 문서와 그렇지 않은 문서(객관적 문서)를 분리해야한다. 본 논문에서는 트위터 문서로 부터 오피니언 문서만을 추출하는 새로운 방법을 제안한다. 트위터 환경에서 오피니언 문서에 대한 분류나 검색의 어려운 점은 충분한 학습 자료가 존재하지 않다는데 있다 이를 위해 제안된 방법에서는 감성 분류를 위해 트위터와 유사한 외부의 정보를 이용하여 기계학습기반 분류 모델을 생성하고, 이를 응용하여 트위터에서의 오피니언 문서 추출에 적용하였다. 또한 실험을 통하여 제안된 방법의 적용 가능성을 평가하였다.

Abstract Recently, as Social Network Services(SNS) are becoming more popular, much research has been doing on analyzing public opinions from SNS. One of the most important tasks for solving such a problem is to separate opinion(subjective) documents from others(e.g. objective documents) in SNS. In this paper, we propose a new method of retrieving the opinion documents from Twitter. The reason why it is not easy to search or classify the opinion documents in Twitter is due to a lack of publicly available Twitter documents for training. To tackle the problem, at first, we build a machine-learned model for sentiment classification using the external documents similar to Twitter, and then modify the model to separate the opinion documents from Twitter. Experimental results show that proposed method can be applied successfully in opinion classification.

Key Words : Opinion Document, Social Network Service, Twitter, Machine Learning, Sentiment Classification

Ⅰ. 서 론

2000년대 이후 SNS(Social Network Service)의 등장은 사용자들의 인터넷 사용 환경에 큰 변화를 가져왔다. SNS가 대중화되기 이전의 사용자는 주로 인터넷 포털 사이트를 이용하여 문서를 검색하거나 온라인 업무를 처리하는 등, 특정인들에 의해 제공되는 정보를 사용하는 수준에 머물렀다. 그러나 SNS를 사용함으로서 개방적인 웹 환경을 기반으로 한 네티즌들의 정보공유와 참여가 가능하게 된 것이다.

정보 제공자가 제한되지 않는 SNS는 그 특성상 기존의 웹문서와는 차별되는 몇 가지 특징을 갖고 있다. 우선 SNS의 문서들은 대부분 단문(short document)으로 구성된다. 예를 들어 트위터(Twitter)에서 하나의 게

*정회원, 한성대학교 컴퓨터공학과
접수일자 : 2013년 8월 26일, 수정완료 : 2013년 9월 25일
게재확정일자 : 2013년 10월 11일

Received: 26 August, 2013 / Revised: 25 September, 2013 / Accepted: 11 October, 2013
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

시글은 140bytes이하로만 허용된다. 또한 많은 문서들이 객관적인(objective) 사실뿐만 아니라 주관적인 의견(subjective opinion)들이 포함되어 있는 등 다양한 종류의 문서들이 혼재한다. 마지막으로 기존의 웹 문서와는 달리 문법적으로 정제되지 않은 문서들이 비율이 매우 높다. 이러한 특징들은 SNS의 문서가 불특정 다수에 의해 대량으로 생성된다는 것에 기인한다. 일례로 트위터의 경우 전 세계적으로 매일 2억 개가 넘은 게시글이 등록되고 있다.

이러한 환경에서 SNS에서의 검색에 대한 관심이 대두되고 있으며, 최근 대부분의 포털에서는 트위터로 대표되는 SNS에 대한 검색 기능도 제공하고 있다. 그러나 기존의 포털에서 제공되는 검색 기능은 사용자의 의도와는 무관하게 단순히 검색어에 대해 최근에 등록된 순서만으로 그 결과를 제공하고 있다. 최근 들어 일부 연구에서 트위터 검색 기법에 대한 연구가 있었지만 아직까지 초보적인 수준을 벗어나지 못하고 있다[1-3].

트위터에서의 검색은 기존의 TF-IDF로 대표되는 웹 검색 방식과는 매우 다르다[4]. 우선 각 문서는 140bytes 이하의 단문이므로 내용(contents)만으로 게시글의 중요도를 평가하기에는 제한적이다. 또한 게시자의 감정 상태나 의견과 같이 오피니언을 제시하는 문서들도 다수를 차지하고 있다. 따라서 트위터에서의 검색은 기존의 웹문서 검색과 같이 실시간 정보 취득의 목적을 위한 검색인지, 아니면 여론 파악을 위해 사용자들의 의견에 대한 검색인지를 구분하는 것이 필요하다. 그러나 현재 대부분의 트위터 검색에 관한 연구는 기존의 웹문서 검색과 같은 상황을 가정하고 있다.

SNS를 이용하여 여론 분석을 위한 오피니언 검색의 핵심 이슈는 주어진 주제에 대해서 긍정(positive)과 부정(negative) 내용을 분리하는데 있다. 이러한 감성 분류(sentiment classification) 기술은 기존의 오피니언 마이닝(opinion mining)에서 이미 많은 연구가 진행되었다[5][6]. 기존의 오피니언 마이닝 연구에서는 주어진 문서가 모두 오피니언 문서라는 가정이 있었다. 그러나 트위터는 오피니언 문서의 비율이 그렇지 않은 문서(객관적인 사실만을 언급한 문서)에 비해 현저히 작은 경우가 흔하다. 따라서 트위터 환경에서 오피니언 문서를 검색하기 위해서는 우선적으로 오피니언 문서만을 분리하는 것이 핵심 기술이라고 볼 수 있다.

본 논문에서는 트위터 환경에서 오피니언 검색을 지원하기 위해 객관적 문서들을 배제하고 오피니언 문서만을 추출하는 방법을 제안한다. 지금까지 텍스트 마이닝 분야에서 기계학습 기반 문서분류는 나이브 베이즈(Naive Bayes) 분류[7], SVM(Support Vector Machine)[7] 등 다양한 기법들이 제안되었다. 이 방법들을 적용하기 위해서는 양질의 학습 문서가 존재해야한다. 그러나 트위터에서는 아직까지 이러한 문서의 확보가 매우 어려운 실정이다. 따라서 본 논문에서는 트위터 문서와 유사한 학습 문서를 이용하여 학습을 한 후 이 정보를 활용하여 트위터에서의 오피니언 문서 추출에 이용하였다. 본 논문에서는 유사 학습문서로서 네이버 영화평을 이용하였다. 네이버 영화평의 경우 140자 이하의 영화평만이 허용되므로 트위터와 문서의 제한 길이가 매우 유사하며, 대부분의 문서가 오피니언 문서로 구성되어 있다. 또한 사용자가 평점을 동시에 부과하게 되어 있어 긍정과 부정 영화평을 쉽게 자동으로 분류해낼 수 있다.

본 논문에서는 특성 선택 기법으로 χ^2-통계량[8]과 KL 거리(Kullback-Leibler divergence)[9]를 이용하였다. 극성(polarity) 분류모델로는 나이브 베이즈 모델과 SVM을 모두 사용하였다. 마지막으로 이렇게 생성된 분류모델들을 수정하여 트위터에서 오피니언 문서를 추출하는데 응용하였다.

제안한 방법을 평가하기 위하여 실험을 실시하였다. 실험은 특성 선택 기법으로 사용된 두가지 기법과 분류를 위해 사용된 두 가지 기법 각각의 조합에 대해 성능을 비교하였다. χ^2-통계량과 나이브 베이즈 모델만을 이용한 일부 성능결과는 [10]에서 이미 제시되었다. 따라서 본 논문에서는 나머지 방법론의 조합을 이용한 확장된 실험 결과를 제시한다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 관련연구에 대해 소개하고, 3장에서는 오피니언 문서 분류 절차에 대해 설명한다. 4장에서는 특징선택 기법을 소개하고 5장에서는 분류 모델에 대해 설명한다. 6장에서는 실험결과를 제시하고 마지막으로 7장에서는 결론을 맺는다.

II. 관련연구

현재까지 트위터에서의 검색을 위한 다양한 방법이 제안되어 왔다[1][2][3][11]. 이러한 연구의 대부분은 중요한

정보로서의 가치가 있는 문서들을 검색하는데 중점을 두고 있다. 즉, 검색을 통하여 새로운 정보를 취득하려는 사용자에 중점을 둔 검색 방식이 대부분이다. 이러한 환경에서 문서 내에 단어의 빈도만으로 관련성을 평가하는 전통적인 TF-IDF 방식은 문제가 있다. 단문으로 구성된 트위터 게시글은 문서 내에 주어진 검색어가 두개이상 존재하는 경우가 매우 드물기 때문이다. 따라서 기존의 트위터 검색 기법에 관한 연구에서도 게시글의 내용보다는 그 이외의 정보를 활용하여 검색에 활용하고 있다. 즉, 문서의 내용보다는 관련된 메타(meta) 정보를 이용할 수밖에 없는데, 각 문서에 대한 메타정보로는 게시자의 팔로워(follower) 수, 재전송(retweet) 빈도, 링크(link) 정보의 포함유무, 멘션(mention) 빈도, 해시태그(hashtag) 등을 예로 들 수 있 있다.

오피니언 문서를 대상으로 한 검색연구는 비교적 최근에 이루어지고 있다[12-15]. 특히 2006년 TREC(Text Retrieval Conference)에서 blog track이 시작된 이후 오피니언 문서 검색에 대한 관심이 높아지고 있어 그 이후에 많은 연구들이 진행되어 왔다. 대표적으로 [12]에서는 토픽(topic) 사전과 감성단어 사전을 별도로 구축하고 검색어와의 연관성을 확률로 각각 계산한 후에 이 값들을 더한 최종 점수로 검색 결과를 제공하는 방법을 제안하였다. [13]에서는 그래프 모델에 기반한 오피니언 문서 검색 기법을 제안하였다. 이 기법에서는 토픽과 감성 단어의 쌍(pair)을 그래프로 표현하고 HITS 알고리즘을 이용하여 검색어와 문서의 연관성을 계산하였다. 그러나 이 연구들에서는 비교적 장문의 문서를 대상으로 하고 있어 단문의 트위터에는 적합한 방법이 될 수 없다.

트위터 문서를 대상으로 한 오피니언 마이닝에 대한 연구도 일부 이루어지고 있다. 대표적으로 [16]에서는 긍정 혹은 부정의 의미를 나타내는 이모티콘(emoticon)을 이용하여 학습을 위한 문서들을 생성하는 기법을 제안하였다. 그러나 한글로 작성된 트위터 문서의 경우 감정을 표현하는 이모티콘을 사용하는 경우가 많지 않아 직접적으로 활용하기에는 한계가 있다. 따라서 본 논문에서는 트위터와 유사한 특성을 지닌 다른 도메인의 문서를 이용하여 분류 모델을 만들고 이를 활용한 트위터에서의 오피니언 문서 추출 방법을 제안한다.

Ⅲ. 오피니언 문서 추출 절차

본 논문에서 수행한 오피니언 문서의 추출 절차는 그림 1과 같다. 우선 네이버 영화평을 이용하여 학습을 위한 데이터를 수집한 후, 평점을 이용하여 긍정과 부정 문서들로 구분한다. 다음 단계로 형태소 분석을 통하여 단어패턴들을 추출한다. 순수한 문서분류에서는 대부분 개별 단어(unigram)만으로도 수행이 가능하지만, 본 논문에서는 unigram, bigram, trigram까지 단어패턴을 추출하였다. trigram까지 추출한 이유는 우선 오피니언 문서에서 '않'과 같은 부정어(negative)를 별도로 하지 않고, 다른 단어와 결합하여 일괄적으로 처리하기 위해서이다. 또 다른 이유는 한글의 경우 하나의 형용사로 의견을 표현하는 경우도 있지만, '쓰레기 영화'와 같이 명사들의 결합으로 긍정이나 부정의 의미를 표현하는 경우도 많기 때문이다.

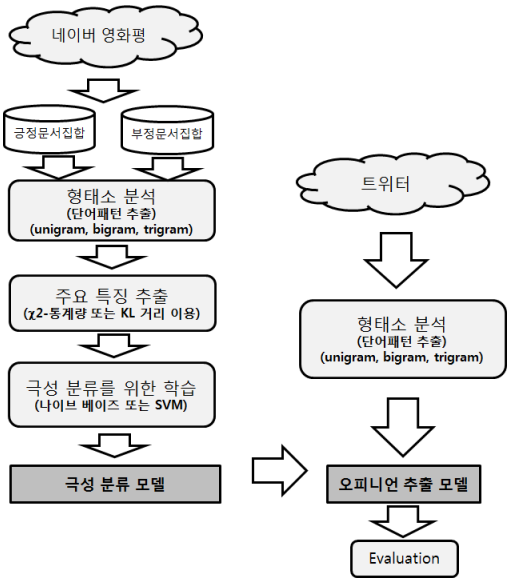


그림 1. 오피니언 문서 추출 절차
Fig. 1. Process of opinion document retrieval

다음 단계로 학습에 사용될 주요 단어패턴들을 선정한다(특징 추출). 일반적으로는 형태소 분석으로 생성된 모든 단어패턴을 특징으로 사용할 수도 있지만, 오피니언 문서에서는 작성자의 감정을 표현하는 특징만을 선정해야 감성분류의 정확도를 높일 수 있다. 따라서 긍정과 부정 문서를 가장 잘 표현하는 특징만을 선

별하는 기법이 필요하다. 본 논문에서는 이를 위해 χ^2-통계량[8]과 KL 거리(Kullback-Leibler divergence)[9]를 이용하였다. 이에 대한 자세한 내용은 다음 장에서 설명한다.

다음으로 기계학습을 통하여 극성(polarity) 분류를 위한 모델을 생성하는데, 본 논문에서는 나이브 베이즈 모델과 SVM을 이용하여 분류모델을 이용하였다. 이렇게 생성된 분류 모델은 극성 분류를 위한 모델이므로 트위터로부터 오피니언 문서를 추출하기 위한 모델로는 직접적으로 사용할 수 없다. 하지만 극성분류를 위한 결정값(decision value)을 이용하면 간접적으로 오피니언 문서와 객관적 문서를 판정하는데 응용할 수 있다. 분류 모델의 결정값은 어떠한 모델을 사용하느냐에 따라 달라질 수 있다. 예를 들어 나이브 베이즈 모델이나 결정 트리(decision tree)의 경우는 각 카테고리에 속할 확률값이 될 수 있고, SVM의 경우 초평면(hyperplane)과의 거리가 될 수 있다. 이에 대한 자세한 내용은 5장에서 설명한다. 마지막으로 테스트 문서를 이용하여 분류의 정확도를 실험한다.

Ⅳ. 특징 선택 기법

오피니언 마이닝에서는 감성 단어가 분류에 큰 역할을 하므로 긍정 및 부정 카테고리에서 추출된 모든 단어패턴을 학습을 위한 특징으로 사용할 수 없다. 또한 본 논문에서는 감성 분류를 위해 특별한 사전을 이용하지 않으므로 각 카테고리를 대표할 수 있는 단어패턴의 선정이 매우 중요하다. 본 논문에서는 각 카테고리에서의 대표적인 단어벡터를 구하기 위한 방법으로 χ^2-통계량과 KL 거리를 이용하였다.

1. χ^2-통계량

χ^2-통계량은 문서분류 분야에서 최적의 특징을 추출하는 데 많이 응용되고 있다[8][17]. χ^2-통계량에서는 모든 특징에 대해 문서가 특정 카테고리의 주제를 표현하는 정도를 정량적으로 평가하여 가장 적합한 특징들을 선택하는데 사용할 수 있다. 주어진 단어패턴 w와 카테고리 c에 대해서 χ^2-통계량 $\chi^2(c, w)$는 w와 c의 관련성 정도를 평가하는 것으로, 이 값이 작으면 서로 독립적이라는 것을 의미하며 반대로 크면 상호 관련성이 크다는 것을 의미한다. $\chi^2(c, w)$는 다음과 같이 계산된다.

$$\chi^2(c, w) = \frac{N\times(a\times d - b\times c)^2}{(a+b)\times(a\times c)\times(b+d)\times(c+d)} \quad (1)$$

where

$$a = DF(w,c) \quad b = DF(w,\bar{c}) \quad c = DF(\bar{w},c) \quad d = DF(\bar{w},\bar{c})$$

여기서 $DF(w, c)$는 w가 포함되는 문서 중에 카테고리 c에 해당에는 문서의 빈도수를 나타내고, $DF(\bar{w}, c)$은 카테고리 c에 포함된 문서 중에 w를 포함하지 않는 문서의 빈도수를 나타낸다. 역으로, $DF(w, \bar{c})$는 c에 포함되지 않으면서 w를 포함한 문서의 빈도수를 나타내며, 마지막으로 $DF(\bar{w}, \bar{c})$는 c에도 포함되지 않고 w도 갖고 있지 않는 문서의 수를 나타낸다. N은 총 문서의 수를 나타낸다. 이 식에 의해서 c와 w가 서로 독립적이면 $\chi^2(c, w)$는 0의 값을 갖게 되며, 반대로 w가 카테고리 c의 주제를 반영하는 단어패턴이면 $\chi^2(c, w)$값은 증가될 것이다.

본 논문에서는 모든 단어패턴에 대해 긍정 및 부정 카테고리에 대한 χ^2-통계량을 계산하여 일정 수준 이상의 값을 갖는 단어패턴만을 학습을 위한 특징으로 선정하였다.

2. KL 거리

KL 거리는 하나의 확률 분포(probability distribution)가 다른 분포와 얼마나 다른가를 정량화한 수치로 정보이론에서 폭 넓게 사용되고 있다. 일반적으로 확률분포 p와 q가 있을 때, 이들 간의 KL 거리는 다음과 같이 정의된다[9].

$$KL(p, q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \quad (2)$$

이 KL 거리함수를 응용하면 문서 d에 대해서 두 카테고리 c_i, c_j의 단어패턴 분포에 대한 KL 거리함수를 다음과 같이 정의할 수 있다[18].

$$KLdist_d(p(W|c_i), p(W|c_j)) = \sum_{w \in W} p(w|c_i) \cdot \log \frac{p(w|c_i)}{p(w|c_j)} \quad (3)$$

여기서 W는 문서 d에 출현하는 모든 단어패턴집합을 의미한다. 이 식에서 KL 거리가 크다는 것은 현재 학습문서의 단어패턴 분포가 문서 d에 출현하는 단어에 대해서는 분별력이 강하다는 것을 의미한다. 즉, 문서 d에 대해서 카테고리에 대한 분별력이 큰 단어분포가 형성되기 위해서는 카테고리에 대한 단어분포의 KL 거리 값이 커야한다. 따라서 모든 학습문서에 대해서 긍정 및 부정 카테고리에 대한 KL 거리을 계산하여 일정 수준 이상의 값을 갖는 문서의 단어패턴만을 학습을 위한 특징으로 선정할 수 있다.

V. 오피니언 분류 및 추출 기법

본 논문에서는 트위터에서의 오피니언 문서 분류를 위해 네이버 영화평을 이용하였다. 이 문서 집합으로부터 형태소분석기를 통해 단어들을 추출하는데, 하나의 문서로 부터 추출된 단어를 이용하여 unigram, bigram, trigram을 구성한 단어패턴들의 집합을 생성하고, 이 패턴들을 이용하여 분류 모델을 생성하였다. 또한 생성된 분류 모델을 이용하여 트위터에서의 오피니언 추출에 활용하였다. 본 논문에서는 이를 위해 나이브 베이즈 분류 기법과 SVM을 이용하였다.

1. 나이브 베이즈 모델

나이브 베이즈 모델은 다음의 식을 이용하여 주어진 문서에 대한 카테고리의 사후확률(posteriori probability)값을 추정함으로써 이루어진다[7, 18].

$$arg\,max_{c_i \in C} Pr(c_j|d_i) = arg\,max_{c_j \in C} Pr(c_j) \cdot \prod_{k=1}^{|d_i|} Pr(w_{ik}|c_j) \quad (4)$$

이 식에서 d_i는 하나의 문서를 나타내며, 단어패턴들의 다중 집합인 $(w_{i1}, w_{i2}, ..., w_{i|d_i|})$로 표현된다. C는 카테고리 집합으로 본 논문에서는 C는 $\{c_P, c_N\}$으로 구성된다. 여기서 c_P와 c_N는 각각 긍정, 부정 카테고리를 나타낸다. 따라서 이 식을 이용하여 트위터 문서 d_i는 사후확률값인 $arg\,max_{c_j \in C} Pr(c_j|d_i)$에 해당하는 카테고리 c_j로 할당된다.

나이브 베이스 분류 모델은 문서의 극성 판별을 위해 사용하였다. 따라서 이 모델은 오피니언 문서와 객관적 문서를 분류하기 위한 방법에 직접적으로 이용할 수 없다. 그러나 나이브 베이즈 모델에서는 주어진 문서에 대해서 각각 긍정과 부정 카테고리에 포함될 확률을 계산하므로, 이 값들을 이용하면 주어진 문서가 어느 정도의 강도(strength)로 오피니언 문서에 가까운지를 추정할 수 있다. 즉, 확률값의 차이가 크면 오피니언 문서로 판정할 수 있고, 그 차이가 없거나 상대적으로 작으면 객관적 문서로 판정할 수 있다. 본 논문에서는 이를 위해 d_i가 각각 c_P와 c_N로 분류될 확률들을 계산하고, 확률값의 차이가 크면 오피니언 문서로 판정하고, 상대적으로 작으면 객관적 문서로 판정하였다. 즉 d_i에 대해서 다음의 계산식을 이용하였다[10].

$$\alpha = \left| Pr(c_P) \cdot \prod_{k=1}^{|d_i|} Pr(w_{ik}|c_P) - Pr(c_N) \cdot \prod_{k=1}^{|d_i|} Pr(w_{ik}|c_N) \right| \quad (5)$$

이 식을 이용하여 α가 일정 기준 값 이상이면 오피니언 문서이며, 반대의 경우 객관적 문서로 분류한다.

2. SVM 모델

SVM은 분류를 위해 개발된 감독형(supervised) 기계학습 기법중 하나로 현재 회귀분석(regression), 분류, 랭킹(ranking) 등에 현재 널리 사용되고 있다. SVM의 개본 개념은 주어진 데이터에 대해서 각 카테고리를 구별하면서 최대의 마진(margin)을 갖는 최적의 초평면을 탐색하는 기법이다[8]. SVM은 $(x_i, y_i)(i = 1, ..., n)$ 형태의 학습 데이터를 가정한다. 여기서 x_i는 분류된 데이터를 나타내며, y_i는 {1, -1}중 하나를 갖는 클래스 라벨(class label)을 나타낸다. 예를 들어 본 논문에서 x_i는 개별 영화평 문서에 대한 단어패턴들을 벡터 형태로 표현한 것이며, y_i에서 +1은 긍정, -1은 부정 영화평을 의미한다. 이들로부터 SVM은 다음 수식이 최소가 되는 매개변수 값을 찾는다.

$$\tau(\mathrm{w}) = \frac{1}{2}\|\mathrm{w}\|^2 \quad (6)$$
$$\text{subject to } y_i((\mathrm{w} \bullet x_i)+b) \geq 1, i=1,...,n$$

이렇게 찾아진 매개변수에 대해서 최종 분류 판정을 위한 결정함수(decision function)는 다음과 같다[19].

$$f(x) = sign(\sum_{i=1}^{n} y_i \alpha_i \bullet (x \bullet x_i) + b) \quad (7)$$
$$\text{where } \alpha_i \text{ is a Lagrange multiplier}$$

식 (7)의 결정함수는 +1 또는 -1의 값으로 결정되므로 이 값에 따라 해당 클래스로 분류된다. 따라서 영화평 문서들에 대해서 긍정 및 부정 학습 문서들을 분류한 후, 각 문서에 대한 단어패턴을 이용하여 극성 분류 모델을 생성할 수 있다.

나이브 베이브 모델과 마찬가지로 이 함수는 극성 분류를 위한 모델이므로 오피니언 문서 추출을 위한 함수로 직접 활용이 불가능하다. 그러나 식 (7)에서 sign 함수 내부의 값은 초평면과 거리가 멀수록 절대값이 더욱 커지게 된다. 따라서 이 식의 절대값이 크면 오피니언 문서일 가능성이 높고, 반대로 0에 가까울수록 객관적 문서에 가깝다고 볼 수 있다. 이를 이용하여 오피니언 문서 추출을 위해 다음의 수식을 활용할 수 있다.

$$\beta = \left| \sum_{i=1}^{n} y_i \alpha_i \bullet (x_i \bullet \bar{d}) + b \right| \quad (8)$$

이 식에서 $\bar{d}$는 문서 d에 나타나는 단어패턴들에 대한 벡터이다. 이 식을 이용하여 β가 일정 기준 값 이상이면 d를 오피니언 문서로 분류하며, 반대의 경우 객관적 문서로 분류할 수 있다.

Ⅵ. 실험평가

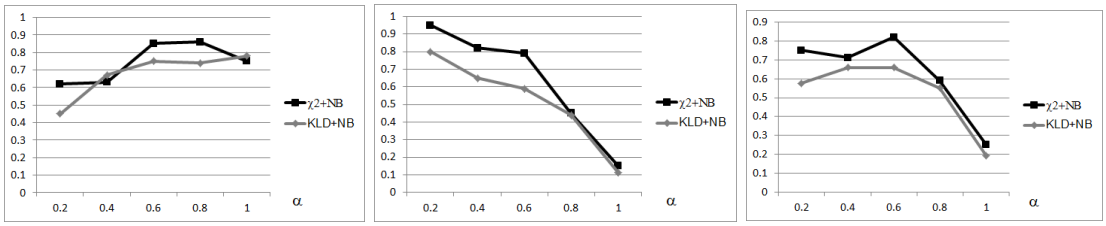
본 논문에서 제안한 오피니언 추출 기법의 적용 가능성을 평가하기 위해서 실험을 실시하였다. 실험을 위해 네이버 영화평 중에서 최근에 흥행에 성공한 영화를 선정하였고, 1부터 5까지의 평점을 부정적인 영화평, 9~10을 긍정적인 영화평으로 간주하였다. 총 30,000여개의 문서를 수집한 후 감성분류 모델을 생성하기 위한 학습을 실시하였다. 트위터에서는 큐로보 사이트(http://www.qrobo.com)로부터 동일한 영화에 대해 관련 문서들을 수집했으며, 이 중에서 오피니언 문서 300개와 객관적 문서 300개를 수작업으로 분류하였다. 특징선택을 위한 χ^2-통계량과 KL 거리에서는 각각 모든 단어패턴 중에서 상위 30%만을 특징으로 최종 선정하였다. 이렇게 선정된 특징들에서 대해서 5장에서 설명한 나이브 베이즈 방법과 SVM 방법을 이용하였는데, 실험은 다음과 같은 총 4가지 조합에 대해서 실시하였다.

- χ^2+NB : χ^2-통계량과 나이브 베이즈
- χ^2+SVM : χ^2-통계량과 SVM
- KLD+NB : KL 거리와 나이브 베이즈
- KLD+SVM : KL 거리와 SVM

실험 결과는 그림 2와 같다. 이 그림에서 (a), (b), (c)는 χ^2+NB와 KLD+NB의 성능 비교를 나타내며, 각각 정확도(precision)와 재현율(recall), 그리고 F-value를 보여준다. 여기서는 공통적으로 나이브 베이즈 방법을 이용했으므로 식 (5)의 α 값 변화에 따른 성능을 측정하였다. 반면에 그림 2의 (d), (e), (f)는 χ^2+SVM와 KLD+SVM의 성능을 나타낸다. 이 경우는 SVM을 이용했으므로, 식 (8)의 β 값에 따른 성능을 측정하였다.

우선 그림 2(a)와 (b)을 보면 공통적으로 α가 증가할수록 정확도는 증가하는 반면 재현율은 반대로 감소하는 경향을 보이고 있다. α가 증가하면 극성의 표현이 강한 문서만을 오피니언 문서로 평가하므로 정확도이 높아지고, 반대로 재현율이 낮아지게 된다. 그림 2(d)와 (e)의 β에 따른 성능도 α와 동일한 현상을 보인다.

다음으로 그림 2(c)를 보면, χ^2+NB가 KLD+NB에 비해 전반적으로 좋은 성능을 보인다. 그림 2(f)에서도 χ^2+SVM이 KLD+SVM에 비해 더 좋은 성능 보이고 있다. 따라서 특징선택에 있어서 χ^2-통계량이 KL 거리에 비해서 더 좋은 성능을 보인다고 볼 수 있다. 본 논문에서 사용한 χ^2-통계량은 긍정과 부정으로 구분된 학습문서집합에서 각 카테고리를 대표하는 단어패턴들을 직접 선택하는 방법을 사용하였다. 반면에 KL 거리에서는 각 학습문서집합을 대표하는 문서들을 선



(a) χ^2+NB와 KLD+NB의 정확도 (b) χ^2+NB와 KLD+NB의 재현율 (c) χ^2+NB와 KLD+NB의 F-value

(a) Precision of χ^2+NB and KLD+NB (b) Recall of χ^2+NB and KLD+NB (c) F-value of χ^2+NB and KLD+NB

(d) χ^2+SVM와 KLD+SVM의 정확도 (e) χ^2+SVM와 KLD+SVM의 재현율 (f) χ^2+SVM와 KLD+SVM의 F-value

(d) Precision of χ^2+SVM and KLD+SVM (e) Recall of χ^2+SVM and KLD+SVM (f) F-value of χ^2+SVM and KLD+SVM

그림 2. 성능실험 결과
Fig. 2. Results of performance experiments

택하고, 선택된 문서에 포함된 단어패턴들을 특징으로 선택하였다. 따라서 KL 거리에 의해 선택된 문서라 할지라도, 해당 문서에 긍정과 부정을 의미하지 않는 중립적인 단어패턴들이 포함될 가능성이 많기 때문에 전반적인 성능이 χ^2-통계량에 비해 떨어진다고 해석할 수 있다. 따라서 KL 거리를 사용한다 할지라도 각 카테고리를 대표하는 문서가 아닌 단어패턴 중심으로 선택하는 기법을 개발한다면 더 우수한 성능을 기대할 수 있을 것으로 판단된다.

나이브 베이즈 방법과 SVM은 α와 β값의 기준이 다르므로 이 값들의 변화에 따른 직접적인 성능 비교는 쉽지 않다. 하지만 그림 2의 (c)와 (f)에서 χ^2-통계량을 사용한 χ^2+NB와 χ^2+SVM를 보면, α가 0.6, β가 1.0 일 때 가장 좋은 F-value를 보이고, 그 값은 각각 0.82, 0.80이다. 또한 같은 그림에서 KL 거리를 사용한 KLD+NB와 KLD+SVM 에서도 α가 0.6, β가 1.0 일 때 가장 좋은 F-value를 보이고 있으며, 그 값은 각각 0.66, 0.67이다. 따라서 가장 좋은 성능을 보인 경우만을 비교해보면 나이브 베이즈 방법과 SVM은 큰 차이를 보이지 않는다고 볼 수 있다.

지금까지의 실험결과를 종합하면 네이버 영화평은 트위터에서의 오피니언 문서 추출을 위한 외부 학습문서로서의 가치가 충분하다고 판단된다. 또한 특징선택에 있어서는 χ^2-통계량이 KL 거리에 비해 더 좋은 성능을 보인 것을 확인할 수 있었다. 마지막으로 오피니언 문서 추출 모델로서는 나이브 베이즈와 SVM은 큰 성능 차이를 보이지 않았다.

Ⅶ. 결 론

본 논문에서는 트위터 문서들 중에서 오피니언 문서만을 추출하는 방법을 제안하였다. 학습 데이터가 충분하지 않은 상황에서 트위터와 유사한 네이버 영화평을 이용하여 감성 분류 모델을 생성한 후 이를 이용하여 오피니언 문서 추출에 적용하였다. 실험 결과 네이버 영화평에서 극성 차이가 일정 수준이상의 경우 트위터에서의 오피니언 문서 추출에 적용될 수 있음을 보였다. 특징추출 방법으로는 χ^2-통계량과 KL 거리를 활용하였으며, 실험 결과 χ^2-통계량이 더 우수한 성능을

보였다. 오피니언 추출 모델에서는 나이브 베이즈 방법과 SVM을 사용하였으며, 실험적으로 큰 성능 차이를 보이지 않았다

본 논문에서 제안된 방법은 양질의 학습문서가 존재하지 않는 트위터 환경을 위해 개발되었다. 트위터는 오피니언 문서뿐만 아니라 객관적 사실을 전달하기 위한 문서, 광고성 문서 등 다양한 종류의 문서들이 혼재한다. 따라서 트위터 문서에 대한 올바른 검색을 지원하기 위해서는 이러한 다양한 문서들에 대한 정확한 분류가 우선되어야하며, 이를 위해서는 트위터를 위한 양질의 학습문서 개발이 필요하다고 하겠다.

References

[1] R. Nagmoti and M. D. Cock, "Ranking Approach for Microblog Search", Proceedings of WI-IAT conference, 2010.

[2] A. Sarma, At. Sarma, S. Gollapudi, and R. Panigrahy, "Ranking Mechanisms in Twitter-like Forums", Proceedings of WSDM conference Feb. 2010.

[3] H. W. Lauw, A. Ntoulas, and K. Kenthapadi, "Estimating the Quality of Postings in the Real-time Web", Proceedings of SSM conference, 2010.

[4] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition), ACM, 2011.

[5] E. Courses and T. Surveys, "Using SentiWordNet for multilingual sentiment analysis", Proceedings of Data Engineering Workshop, 2008.

[6] Q. Miao, Q. Li, and R. Dai, "A sentiment mining and retrieval system", Expert Systems with Applications, Vol.36, pp. 7192-7198, 2009.

[7] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison-Sesley, 2006

[8] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proceedings of ECML/PKDD-2006 International Workshop on Knowledge Discovery from Data Streams, 2006.

[9] T. M. Cover and J. A. Thomas, Elements of Information Theory, Wiley, New York, 1991.

[10] J. Chang, S. Lee, and J. Han, "Machine-Learned Classification Technique for Opinion Documents Retrieval in Social Network Services", Proceedings of 2013 Korea Computer Congress, 2013.

[11] J. Chang, "An Evaluation of Twitter Ranking Using the Retweet Information", Journal of Korea Society for E-Business Studies, Vol. 17, No. 2, 2012.

[12] X. Huang and W. B. Crott, "A Unified Relevance Model for Opinion Retrieval", Proceedings of CIKM '09, 2009.

[13] B. Li, L. Zhou, Shi Feng, and K. Wong, "An efficient approach for sentence-based opinion retrieval", Proceedings of 48th Annual Meeting of the Association for Computational Linguistics, pp. 1367-1375, 2010.

[14] W. Zhang, C. Yu, and W. Meng, "Opinion Retrieval from Blogs", Proceedings of CIKM '07, 2007.

[15] J. Chang, "Efficient Retrieval of Short Opinion Documents Using Learning to Rank", Journal of the Institute of Internet, Broadcasting and Communication, Vol. 13, No. 4, Aug., 2013.

[16] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision", CS224N Project Report, Stanford, 2009.

[17] H. Kim, and J. Chang, "Improving Naive Bayes Text Classifiers with Incremental Feature Weighting", Journal of Korea Information Processing Society, Vol. 15-B, No. 5, 2008.

[18] J. Chang, and H. Kim, "Accelerating the EM Algorithm through Selective Sampling for Naive Bayes Text Classifier", Journal of Korea Information Processing Society, Vol. 13-D, No. 3, 2006.

[19] T. Joachims, “Making large-Scale SVM Learning Practical. Advances in Kernel Methods”, Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[20] M. Hwang, D. Choi, and P. Kim “A Context Information Extraction Method according to Subject for Semantic Text Processing”, Journal of Korean Institute of Information Technology, vol. 8, No. 11, pp. 197-204, 2010.

[21] J. Shim, H. C. Lee, “The Development of Automatic Ontology Generation System Using Extended Search Keywords” Journal of the Korea Academia-Industrial Cooperation Society, Vol. 11, no. 6, 2009.

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.

저자 소개

장 재 영(정회원)

- 1992년 : 서울대학교 계산통계학과 (이학사)
- 1994년 : 서울대학교 계산통계학과 (이학석사)
- 1999년 : 서울대학교 계산통계학과 (이학박사)
- 2000년~현재 : 한성대학교 컴퓨터공학과 교수

<주관심분야 : 데이터베이스, 정보검색, 데이터마이닝>