http://dx.doi.org/10.7236/JIIBC.2013.13.4.117
JIIBC 2013-4-16

기계학습을 이용한 단문 오피니언 문서의 효율적 검색 기법

Efficient Retrieval of Short Opinion Documents Using Learning to Rank

장재영*

Jae-Young Chang

요 약 최근 들어 트위터나 페이스북과 같은 SNS가 대중화되면서, 오피니언 마이닝에 관한 연구가 활발히 진행되고 있다. 그러나 현재의 오피니언 마이닝 연구는 대부분 감성분류나 특징선택 방법에 중점을 두고 있으며, 오피니언 문서의 검색에 관한 연구는 아직 미진한 실정이다. 본 논문에서는 단문으로 구성된 오피니언 문서로부터 사용자가 원하는 문서들을 효율적으로 검색하는 기법을 제안한다. 제안된 방법에서는 기존의 감성분류 방법을 활용함과 동시에 문서의 질적 평가를 위해 여러 가지 특징들을 적용한다. 검색 모델을 생성하기 위해 기계학습 기반 랭킹 기법을 활용하며, 감성 분류 모델을 기계학습 랭킹 모델에 통합하는 방법을 사용한다. 또한 실험을 통하여 제안된 방법이 오피니언 검색에 효율적으로 적용될 수 있음을 보여준다.

Abstract Recently, as Social Network Services(SNS), such as Twitter, Facebook, are becoming more popular, much research has been doing on opinion mining. However, current related researches are mostly focused on sentiment classification or feature selection, but there were few studies about opinion document retrieval. In this paper, we propose a new retrieval method of short opinion documents. Proposed method utilizes previous sentiment classification methodology, and applies several features of documents for evaluating the quality of the opinion documents. For generating the retrieval model, we adopt Learning-to-rank technique and integrate sentiment classification model to Learning-to-rank. Experimental results show that proposed method can be applied successfully in opinion search.

Key Words : Opinion Document, Search, Learning to Rank, Sentiment Classification

Ⅰ. 서 론

최근 들어 스마트폰으로 대표되는 모바일 기기가 점차 대중화되면서 특정장소에 국한되지 않고 인터넷을 접근할 수 있는 환경이 마련되고 있다. 사용자들은 언제 어디서든지 인터넷을 통해서 새로운 정보를 습득할 수 있을 뿐만 아니라 정보의 제공자 역할도 하고 있다. 더구나 트위터(Twitter)나 페이스북(Facebook)과 같은 SNS(Social Network Service)의 등장은 이러한 정보교류의 폭발적 성장에 큰 역할을 담당하고 있다.

*정회원, 한성대학교 컴퓨터공학과
접수일자 : 2013년 5월 14일, 수정완료 : 2013년 6월 27일
게재확정일자 : 2013년 8월 16일

Received: 14 May, 2013 / Revised: 27 June, 2013 / Accepted: 16 August, 2013
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

정보 제공자가 제한되지 않는 SNS는 그 특성상 기존의 웹문서와는 차별되는 몇 가지 특징을 갖고 있다. 우선 SNS의 문서들은 비교적 단문(short document)으로 구성된다. 또한 많은 문서들이 객관적인(objective) 내용뿐만 아니라 주관적인 의견(subjective opinion)들이 많이 포함되어 있다. 마지막으로 기존의 웹 문서와는 달리 문법적으로 정제되지 않은 문서들이 다량으로 발생하고 있다. 이러한 환경에서 최근의 대부분의 포털에서는 SNS에 대한 검색 기능도 제공하고 있다. 그러나 기존의 포털에서 제공되는 검색 기능은 사용자의 의도와는 무관하게 단순히 검색어(query)에 대해 최근에 등록된 순서만으로 그 결과를 제공하고 있다. 앞서 언급한 바와 같이 SNS에서의 문서들은 객관적인 사실뿐만 아니라 주관적인 의견이 포함된 문서들도 존재하고 있으며, 문서의 질적인 측면에서도 매우 다양한 스펙트럼을 형성하고 있다. 따라서 TF-IDF로 대표되는 기존의 웹 검색 방식만으로는 사용자의 의도를 반영한 검색에는 한계가 있다[4][5][6][7][8].

주관적 문서의 처리 기법에 대한 대표적인 연구 분야로는 오피니언 마이닝(opinion mining)을 들 수 있다[2][9][10][11][12][13]. 하지만 현재의 오피니언 마이닝에 대한 연구는 긍정/부정문서들을 분류하는 기술에 초점이 맞추어져 있고, 오피니언 문서를 질적으로 평가하여 검색에 활용하는 기술은 거의 연구가 되고 있지 않고 있다. 따라서 SNS와 같이 오피니언 문서가 대용량으로 발생되는 현재의 상황에서 사용자가 원하는 오피니언 문서를 검색하기 위해서는 감성 분류(sentiment classification) 기술뿐만 아니라 오피니언 문서의 특성에 맞는 검색 기술에 대한 연구가 필수적이라 하겠다.

본 논문에서는 이러한 환경에서 단문으로 구성된 오피니언 문서로부터 사용자가 원하는 문서들을 효율적으로 추출하는 검색 기법을 제안한다. 오피니언 문서의 검색을 위해서는 우선 긍정(positive)과 부정(negative) 문서를 분류하는 것이 가장 중요하며, 동시에 사용자의 검색의도와 부합되도록 문서에 대한 질적 평가(quality evaluation)도 필요하다. 이를 위해 제안된 방법에서는 기존의 감성분류 방법[2]을 활용함과 동시에, 문서의 질적 평가를 위해 오피니언 문서 검색을 위한 여러 가지 특징(feature)들을 적용한다.

[2]에서 제안된 감성 분류기법은 문법적 요소를 최대한 배제하고 단어패턴의 빈도만을 고려하였다. 이 기법은 비교적 단순하게 감성분류를 할 수 있을 뿐만 아니라 기존의 확률을 기반으로 한 문서 분류 방법을 그대로 적용한 것으로 본 논문이 제안하는 오피니언 검색에서 하나의 특징으로 활용될 수 있다. 문서의 질적 평가를 위해서는 문서 길이, 문법의 구성형태와 같은 외형적 요소뿐만 아니라 문서의 전문성 등을 평가하기 위한 내용적 요소들을 고려하였다. 검색 모델에 있어서 기존의 오피니언 문서 검색에서는 감성분류 기술과 검색기술을 분리하여 적용하였으나 본 논문에서는 기계학습 랭킹(learning to rank)을 활용하여 이들을 하나의 프레임워크로 통합한 모델을 사용하였다.

본 논문에서는 제안된 검색 기술의 평가를 위해 실험을 실시하였다. 실험은 네이버 영화평을 대상으로 하였다. 네이버 영화평은 140자 이하의 단문으로 구성된 오피니언 문서로 한글로 구성된 오피니언 문서의 분류 실험에 가장 많이 활용되는 데이터 중의 하나이다. 이 문서로부터 본 논문이 제안한 검색 기법을 적용하여 감성 분류의 정확도뿐만 아니라 검색 정확도도 측정하여 오피니언 문서에 대한 검색 기술로의 적용 가능성을 평가하였다.

II. 관련연구

온라인 문서 검색에서 가장 전통적이며 많이 응용된 방법이 바로 TF-IDF이다[4][5]. 이 기법은 문서 내에 단어의 출현 빈도로 검색어와 문서와의 연관성을 정량적으로 평가하여 검색 결과에 반영하게 된다. 웹 검색 분야에서 가장 주목받는 기법으로는 PageRank[17]를 들 수 있다. 이 방법은 월드 와이드 웹과 같은 하이퍼링크(hyperlink) 구조를 가지는 문서의 상대적 중요도에 따라 가중치를 부여하여 검색어와 문서와의 연관도를 계산한다. 이외에서도 이러한 방법들을 응용한 다양한 검색 기술들이 제안되어 왔고 실제 많은 검색 시스템에서 활용되고 있다.

그러나 SNS와 같이 단문으로 구성된 문서에서의 검색은 내용 및 형식면에서 기존 웹문서와 많은 차이점이 존재하므로 기존의 검색 기법을 직접적으로 적용하기에는 한계가 있다. 트위터는 일반 문서와는 달리 재전송(retweet), 멘션(mention), 팔로워(follower), URL의 포함 여부 등 문서들의 특징을 정의할 수 있는 다양한 요소들이 존재한다. 따라서 문서의 내용보다는 이러한 특징들을 활용한 검색 연구가 대부분을 차지하고 있다[6][7][8].

오피니언 문서를 대상으로 한 검색연구는 비교적 최

근에 이루어지고 있다[18][19][20]. 특히 2006년 TREC(Text Retrieval Conference)에서 blog track이 시작된 이후 오피니언 문서 검색에 대한 관심이 높아지고 있어 그 이후에 많은 연구들이 진행되어 왔다. 오피니언 문서 검색의 핵심은 검색어와 문서와의 연관도를 평가함과 동시에 감성분류를 위한 평가도 동시에 실시해야한다는 점이다. 기존의 오피니언 문서 검색에서는 대부분 이러한 작업을 분리해서 진행하고 있다.

지금까지 언급한 오피니언 문서 검색 기법들은 대부분 오피니언 문서(주관적 문서)와 객관적 문서를 판별하는 것이 주된 관심이었으며, 오피니언 문서 중에서 어떤 문서가 사용자의 검색의도와 얼마나 잘 부합하는 가는 큰 관심을 두지 않았다. 또한 대부분 단문이 아닌 비교적 큰 문서를 대상으로 하였다. 반면에 본 논문에서는 SNS와 같이 단문으로 구성된 문서만을 대상으로 하였으며, 주어진 문서가 모두 오피니언 문서라는 가정 하에 사용자의 검색어에 대해서 상대적으로 잘 쓰인 오피니언 문서를 검색하는 데 초점을 두었다. 또한 감성분류 모델을 기계학습기반 검색의 하나의 특징으로 취급함으로써 통합된 오피니언 검색 모델 생성 기법을 제안하였다.

Ⅲ. 단문 오피니언 문서의 특징

개방적인 웹 환경을 기반으로 네티즌들의 정보공유와 참여가 가능하게 된 현재의 환경에서 불특정 다수에 의해 제공되는 문서들은 기존의 웹 문서와 다른 몇 가지 특징을 갖고 있다. SNS나 뉴스기사의 댓글들에서 보는 바와 같이 대부분의 문서들이 일정 길이 이하의 단문으로 구성된다. 또한 문서 제공자의 수와 절대량이 기존의 웹 문서에 비해 매우 방대하고 익명성을 보장하는 경우도 흔하다. 따라서 이러한 문서들은 새로운 정보를 제공하기보다 자신의 의견을 피력하는 오피니언 문서가 주를 이룬다. 따라서 이러한 문서들에 대한 검색은 기존의 방법을 벗어나 그 특성에 맞게 변화해야한다. 본 논문에서 제안하는 검색 방식도 이러한 환경을 반영하여 단문의 오피니언 문서만을 검색 대상으로 한다. 예를 들어 표 1은 네이버 영화평에 대한 문서의 예이다. 네이버 영화평은 익명성을 보장하며 한글 기준 140자 이하로만 작성된다. 이 예는 최근에 개봉한 한 영화에 대한 평가들로 크게 보면 긍정적인 영화평과 부정적인 영화평으로 나눌 수 있다. 또한 각 그룹에서 이 문서를 검색한 사용자에게 도움이 될 만한 좋은 문서들이 있는 반면, 자신의 감정만을 두서없이 표현하여 이 글을 읽는 사용자에게 전혀 도움을 되지 않는 문서들도 존재한다. 사용자가 영화제목을 검색하였을 경우 긍정 혹은 부정적인 영화평을 자동분류하는 것뿐만 아니라 상대적으로 잘 쓰인 문서를 우선적으로 검색결과로 제공해야한다. 이를 위해서는 잘 쓰인 문서 – 본 논문에 서는 영화평만을 가정한다. – 에 대한 정의가 필요하다. 본 논문에서는 이러한 문서를 다음과 같은 기준으로 정의하였다.

표 1. 영화평의 분류 예시
Table 1. An Example of Movie Reviews Classification

	긍정영화평	부정영화평
좋은문서	코믹하면서도 감동적인 스토리가 가슴을 찡하게 만들어눈물샘을 자극하는 영화다. 류승룡 연기력최고였고... 딸 예승이 갈소원도 이 영화에 재미와 감동에 한몫해 정말로 감동적인 영화였다.	눈물은 납니다. 그런데 이게 정말 감동을 받아서 흘리는 눈물이 아니라 억지로 눈물을 쥐어짜게 만들어놨어요. 배우들 연기는 좋지만 영화 자체는 실망이에요. 내용도 진부하고 개연성도 없고요.
좋지않은문서	이 영화는 어굴하고 정말 이런이 생겨서 아픈이 생겨는지... 보는 사람다 아빠는 아픈고 딸바보라 다른사람은 무시해다는 것이 재일 중한것 같다. 보로 갔다며 생각과 마음을 비우고 갔다며 최고 재미있게 볼수있다	이런 고딩이 찍은 과제발표 수준의 영화가 관객 처울린다고 천만이라니 정말 할말이 없다..정말 우리나라 군중심리는 알아줘야해..개인적으로 최고의 쓰레기영화라 평하고 싶다.

- **감정의 표현**

오피니언 문서의 검색에서는 중립적인 표현 보다는 감정의 표현이 명확한 문서를 우선적으로 검색하는 것이 사용자의 검색 의도에 부합한다.

- **풍부한 내용**

간결한 표현 보다는 비교적 내용이 풍부하고 다양할수록 좋은 문서로 평가할 수 있다.

- **문법적 구성**

자유도가 높은 한글 문서의 특성상 맞춤법이나 문법에 벗어난 표현이 많을 수 있다. 그러나 비교적 문법적 표현을 준수한 문서가 성의 있게 작성된 오피니언 문서

로 볼 수 있으며, 이러한 문서위주로 검색결과를 제공해야한다.

• **전문성**

오피니언 문서는 전체적인 단순한 감정 표현보다는 오피니언의 대상되는 객체의 세부 특징들에 대해 자세히 평가한 것이 좋은 문서로 볼 수 있다. 예를 들어 영화의 경우 세부 특징으로는 연기력, 영상, 배우, 감독, 편집 등을 들 수 있다.

표 1에서 보는 바와 같이 잘 쓰인 문서들은 그렇지 않은 문서에 비해 위에서 나열한 기준들을 상대적으로 충족한다고 볼 수 있다. 따라서 본 논문에서는 이러한 기준에 부합한 특징들을 개발하고 이를 이용한 검색 기법을 제안한다.

Ⅳ. 오피니언 문서 검색기법

문서 검색을 위해서는 문서를 평가하기 위한 요소(특징)들을 결정하고 이들을 이용한 검색 모델을 개발해야한다. 우선 문서의 특징들은 3장에서 제시한 좋은 오피니언 문서에 대한 평가 기준에 부합도록 개발되어야한다. 본 논문에서는 기계학습을 이용한 평가 방식을 사용하였다. 본 장에서는 우선 문서를 평가하기 위한 특징들을 정의하고, 기계학습 기반 검색 모델을 설명한다.

1. 오피니언 검색 모델을 위한 특징들

3장에서는 감정의 표현, 풍부한 내용, 문법적 구성, 전문성을 오피니언 문서의 평가 기준으로 정의하였다. 이를 반영하기 위한 특징들은 다음과 같다.

가. 감정의 극성

감정의 극성(polarity)을 판별하고 그 정도를 정량화하기 위해 본 논문에서는 [2]에서 제안한 확률적 접근기법을 사용한다. 이 방법은 단문으로 구성된 오피니언 문서의 극성 판별에 적합한 방법으로 볼 수 있다. 우선 수집된 전체 문서의 집합 중에서 긍정문서 집합을 D_P, 부정문서 집합을 D_N이라 하자. 또한 각 문서로부터 형태소분석기를 통해 단어들을 추출하는데, 하나의 문서로 부터 추출된 단어를 이용하여 unigram, bigram, trigram을 구성한 단어패턴들의 집합을 생성할 수 있다. 이와 같이 문서 d로부터 생성된 단어패턴 집합을 $\overline{d}$로 정의한다. 이와 같은 과정을 통해 D_P와 D_N은 각각 다음과 같이 단어패턴들의 집합인 $\overline{D_P}$와 $\overline{D_N}$으로 재정의할 수 있다.

$$\overline{D_P} = \bigcup_{d \in D_P} \overline{d} \quad \overline{D_N} = \bigcup_{d \in D_N} \overline{d} \tag{1}$$

각 단어패턴 w에 대해서, 이 패턴이 $\overline{D_P}$와 $\overline{D_N}$에 각각 출현한 빈도수를 계산할 필요가 있다. 이를 위해 다음과 같이 $f_P(w)$와 $f_N(w)$를 정의한다.

$$f_P(w) = \overline{D_P}\text{에서 } w\text{의 출현 빈도} \tag{2}$$
$$f_N(w) = \overline{D_N}\text{에서 } w\text{의 출현 빈도}$$

다음으로 각 단어 패턴에 대해 긍정과 부정 중 어느 방향에 가까운지에 대한 정도를 정량적인 확률 값으로 계산해야한다. 정량적 수치를 부여하기 위한 가장 간단한 방법은 조건부 확률을 이용하는 것으로, 단어패턴 w가 긍정 및 부정 단어패턴일 확률은 각각 다음과 같이 정의할 수 있다.

$$p(\overline{D_P}|w) = \frac{p(w|\overline{D_P})\,p(\overline{D_P})}{p(w)} \tag{3}$$

$$p(\overline{D_N}|w) = \frac{p(w|\overline{D_N})\,p(\overline{D_N})}{p(w)} \tag{4}$$

이 식에서 $p(\overline{D_P}|w)$는 w가 긍정 패턴일 확률을 나타내며, 반대로 $p(\overline{D_N}|w)$는 w가 부정 패턴일 확률을 나타낸다. 이 식을 활용하여 주어진 문서가 긍정적인 문서인지 아니면 부정적인 문서인지 판별해야하는데, 이를 해결하기 위해 식 (3)과 (4)에서 계산된 값과 단어들의 출현 빈도를 이용하여 단어의 극성을 계산한다. 주어진 문서 d에 대해 극성을 판단하기 위한 스코어 함수는 다음과 같다.

$$Pscore(d) = \sum_{w \in \overline{d}} (f_P(w) \times p(\overline{D_P}|w)) \tag{5}$$

$$Nscore(d) = \sum_{w \in \overline{d}} (f_N(w) \times p(\overline{D_N}|w)) \tag{6}$$

이 식을 이용하여 문서 d의 $Pscore$ 값과 $Nscore$ 값을 비교하여 더 높은 값으로 d의 극성을 판단할 수 있다. 마지막으로 문서 검색을 위한 특징 값을 위해 감정의 극성 정도를 정량적으로 표현해야하는데 이 문제는 단순히 $Pscore$ 와 $Nscore$의 차이로 알 수 있다. 즉, 이 값들의 차가 클수록 감정의 표현이 한 방향을 분명히 표현된다고 볼 수 있다. 따라서 문서 d에 대한 최종적인 감정의 극성 정도는 다음의 식으로 계산한다.

$$Pscore(d) - Nscore(d) \quad (7)$$

이 식에 의해 긍정문서의 경우 양수의 값을 갖고 부정문서의 경우 음수의 값을 갖는다. 하지만 이 식 (7)을 그대로 이용할 경우에는 긍정과 부정문서모두에 나타나는 중립적인 단어 패턴들에 영향을 받을 가능성이 크다. 따라서 [2]에서 제시한 바와 같이 단어패턴 w가 긍정과 부정 패턴에 모두 나타나는 경우 긍정 혹은 부정 패턴에 속할 확률의 차이가 큰 것만을 선택적으로 이용하는 것이 가장 적합하다. 이 확률의 차이는 다음의 식에 의해 계산할 수 있다.

$$\alpha = |p(\overline{D_P}|w) - p(\overline{D_N}|w)| \quad (8)$$

여기서 계산된 α 값에 대해서 이 값 이상인 경우만을 대상으로 식 (7)을 적용하여 극성을 분류할 수 있다.

나. 문서의 크기

네이버 영화평의 경우 문서의 길이는 최대 140자 이하이다. 그러나 대부분의 게시글들은 이 보다 작은 크기로 표현되고 있다. 본 논문에서 수집한 문서들의 경우 평균 게시글의 크기는 40자 이하이고, 30자 이하의 영화평이 50%이상이다. 물론 큰 문서일수록 좋은 문서라는 보장은 없지만, 지나치게 작은 문서는 좋은 오피니언 문서로 평가하기 어려운 것이 사실이다. 예를 들어 “연기자는 명품 내용은 평범”, “최고...눈물난 작품”, “망설이지 마시고 꼭 보세요..강추입니다”와 같은 문서들은 오피니언을 검색하는 사용자에게 큰 도움이 되지 않는다. 따라서 본 논문에서는 문서의 길이가 길수록 좋은 오피니언 문서라는 가정 하에 문서의 크기(bytes 수)를 하나의 특징 값으로 정의하였다.

다. 문법적 완성도

불특정 다수의 사용자에 의해 작성된 단문은 한글의 문법적 규칙을 어기는 경우가 매우 흔하다. 게시글들을 검색할 때 표 1에서 보는 바와 같이 문법적 규칙을 잘 지킨 문서가 그렇지 못한 문서보다 우선적으로 검색할 필요가 있다. 본 논문에서는 이를 반영하기 위해 한글 형태소 분석기의 결과로서 한글 문법의 품사(Part of Speech) 중에서 분석이 불가능한 품사를 제외한 나머지 품사의 비율을 하나의 특징으로 정의하였다. 즉 문법적 완성도에 대한 특징 값은 다음과 같이 정의된다.

$$\frac{\sum_{PoS(t) \neq NA} |t|}{\| t \|} \quad (9)$$

이 식에서 $\| t \|$ 는 하나의 문서에서 추출된 형태소들의 총 수를 나타낸다. $PoS(t)$는 형태소 t의 품사를 나타내며, NA는 분석불능인 품사 종류를 의미한다.

라. 특성 표현의 다양성

오피니언 문서에 대해 감정 표현을 할 때 대상 객체에 대한 직접적인 표현보다는 객체의 구체적인 특성(feature)으로 세분화하여 표현하는 것이 보다 전문적인 오피니언이라고 볼 수 있다. - 여기서 특성이란 오피니언 대상 도메인의 특징을 말한다. 본 논문에서는 검색 인자로서의 특징과 그 의미를 구분하기 위해 특성이란 용어를 사용한다. - 예를 들어 상품의 경우는 가격, 디자인, 품질 등이 특성이 될 수 있으며, 영화는 연기, 영상, 감독, 편집 등을 특성으로 들 수 있다. 따라서 문서 내에 특성들에 대한 언급이 많을수록 좋은 오피니언 문서로 평가할 수 있다. 이를 평가하기 위한 수식은 다음과 같다.

$$\sum_{t \in f} |t| \quad (10)$$

이 식에서 t는 문서의 각 형태소를 나타내며, f는 해당 오피니언 도메인의 특성 집합을 의미한다. 따라서 문서에서 오피니언 특성들의 언급 빈도를 하나의 특징으로 정의하였다.

마. 대표 단어와의 유사도

오피니언 문서는 크게 긍정 혹은 부정문서로 나눌 수

있고, 각 카테고리 내에서 문서의 질에 따라 여러 단계로 분류할 수 있다. 예를 들어 표 1에서는 문서의 질을 두 개의 세부적인 카테고리로 분류하였다. 이와 같이 오피니언 문서는 여러 종류의 카테고리로 나눌 수 있는데, 기계학습을 이용할 경우에는 각 카테고리에 학습 문서들을 배정하고 이를 이용하여 그 특징들을 정의하게 된다. 전통적인 문서분류를 위한 기계학습에서는 각 카테고리에 포함된 문서들의 단어벡터(bag of words)를 가장 큰 특징으로 가정하고 있다. 이를 위해 본 논문에서도 긍정과 부정으로 분류된 각 카테고리 중에서 최상위 카테고리에 포함된 문서의 대표적인 단어벡터를 추출한 후, 이 벡터와 각 문서에 포함된 단어벡터와의 유사도(similarity)를 계산하여 그 값을 주어진 문서의 특징 값을 정의하였다.

각 카테고리에서의 대표적인 단어벡터를 구하기 위한 방법으로는 χ^2-통계량을 이용하였다. χ^2-통계량은 문서분류에 대한 연구에서 일정 개수의 최적의 특징을 추출하는 데 폭 넓게 응용되고 있다[1][21]. 주어진 단어 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.

본 논문에서는 이 성질을 이용하여 긍정과 부정문서 각각에 대해서 질적인 측면에서 최상위 카테고리에 포함된 문서들의 각 단어들에 대해서 χ^2-통계량 값을 모두 계산하고 상위 값을 갖는 단어들을 선택한다. 그런 다음 각 문서의 단어 벡터와 선택된 단어와의 유사도를 계산하였다. 유사도는 다음과 같이 코사인 유사도(cosine similarity)를 이용하였다.

$$sim(d,d_\lambda) = \frac{\sum_{k=1}^{n} w_k \cdot w_{\lambda k}}{\sqrt{\sum_{k=1}^{n} w_k^2 \cdot \sum_{k=1}^{n} w_{\lambda k}^2}} \qquad (11)$$

이 식에서 $d = (w_1,...,w_n)$ 는 각 문서의 단어벡터이며, $d_\lambda = (w_{\lambda 1},...,w_{\lambda n})$은 χ^2-통계량 값 중에서 상위 값을 갖는 단어들의 벡터이다. 여기서 중요한 것은 긍정문서의 카테고리들 중에서 최상위 카테고리에 대해서 d_λ를 생성하고, 반대로 부정문서의 카테고리들 중에서도 최상위의 카테고리에 대해 d_λ를 생성해야한다. 그 다음 각각에 대해서 유사도를 계산해야한다. 따라서 유사도 측면에서는 두 개의 특징 값을 갖게 된다.

2. 기계학습을 이용한 검색 모델

오피니언 검색을 위해서는 검색어에 대해 문서들의 관련 점수를 부여하는 랭킹 기법이 필요하다. 본 논문에서 검색어에 대한 모델은 $(q, polarity)$와 같은 형태를 가정한다. 여기서 q는 검색어를 나타내며, $polarity$는 $\{P, N, PN\}$중의 하나로 가정한다. 여기서 $polarity$가 P이면 긍정문서들을 우선적으로 검색하는 것이고, N이면 반대로 부정문서들을 우선적으로 검색한다. PN일 경우에는 감정의 극성에 관계없이 문서의 질적인 측면만 고려하여 검색한다. 따라서 이 방식으로 검색할 경우에는 하나의 랭킹 모델로는 학습이 불가능하고 $polarity$에 따라서 별도의 모델이 필요하다.

본 논문은 3장에서 나열한 특징들을 이용하여 기계학습을 이용한 랭킹 방법을 사용한다. 기계학습 랭킹방법으로는 rankSVM을 사용하였다. 이 방법은 SVM을 랭킹 문제에 응용한 pairwise 기법 중 하나로, 기계학습 랭킹에 광범위하게 사용되고 있다. 이 방법은 학습할 문서들을 특징 벡터로 표현한 후, 각 문서에 대한 특징벡터의 쌍인 d_i와 d_j에 대해서, d_i가 d_j에 비해 우선적으로 검색되어야한다면, 다음을 만족하는 최적의 함수 f를 생성하는 것이다.

$$f(d_i) - f(d_j) = w^T d_i - w^T d_j = w^T(d_i - d_j) > 0 \qquad (12)$$

여기서 w는 SVM에서 초평면(hyperplane)의 법선 벡터(normal vector)를 나타낸다. 이와 같은 방법으로 함수 f가 생성되었다면, 주어진 검색어에 대해서 각 문서에 대해 함수를 적용하고, 그 값을 이용하여 문서들에 대한 랭킹 결과를 생성하게 된다.

V. 실험평가

1. 실험 방법

본 논문이 제안한 검색 방식의 성능을 평가하기 위해서 실험을 실시하였다. 실험은 네이버 영화평에서 가장 최근에 흥행에 성공한 영화를 선정하였다. 이 영화에 대

해 평점을 기준으로 1부터 5까지를 부정적인 영화평으로 간주하고 9~10을 긍정적인 영화평으로 간주하였다. 이 문서들을 바탕으로 우선 4.1절에서 설명한 극성 특징에 대한 점수를 계산하기 위한 모델을 생성하였다. 극성 점수를 계산하기 위해 수집된 문서의 총 수는 약 30,000여 개이다.

네이버 영화평은 극성의 분류를 위해서는 평점이라는 객관적인 점수를 이용할 수 있지만 문서의 질을 객관적으로 평가할 수 있는 기준은 존재하지 않는다. 따라서 학습 문서를 위해 수작업으로 문서들을 분류하였다. 이를 위해 수집된 문서 중에서 극성 점수를 바탕으로 긍정과 부정 영화평을 임의로 600개씩 선정하고, 각각에 대해 문서의 질적 측면에서 'best', 'good', 'fair', 'bad'로 나누었다. 표 2는 분류된 학습문서들의 종류와 문서 수를 나타낸다.

표 2. 학습 문서의 수
Table 2. The Number of Documents for Learning

quality / *polarity*	best	good	fair	bad	계
긍정문서	224	184	74	118	600
부정문서	198	162	120	120	600

이 학습 문서를 대상으로 4장에서 선정한 각 특징들에 대한 값을 부여하였다. 극성의 정도는 이미 앞에서 설명을 하였고, 문서의 길이와 문법적 완성도도 정량적 측정이 가능하다. 4.1절에서 설명한 특성 표현의 다양성을 위해 영화 도메인에 대한 특성들을 정의하였다. 본 실험에서 정의한 대표적인 특성들은 다음과 같다.

연기, 영상, 감독, 편집, 배우(배우명 포함), 배역, 조명, 의상, 분장, 세트, 연출, 특수효과, 음악, 녹음, 장면, 촬영, 표현, 카메라, 시나리오, 내용, 스토리 등

최적 단어벡터와의 유사도는 긍정 및 부정으로 분류된 문서 중에서 'best'로 분류된 문서들에 대해 각각 χ^2-통계량을 이용하여 50개씩의 단어를 선정하고, 이들과 학습 문서의 단어 벡터와의 유사도를 특징 값으로 정의하였다.

마지막으로 rankSVM을 적용하기 위해서는 각 학습 문서그룹에 대해 검색어와 문서들과의 관련 점수(relevance score)를 부여해야한다. 본 논문에서는 세 가지의 검색 방식을 위해 표 3과 같이 점수를 부여하였다. 이 표에서 보는 바와 같이 $polarity$가 P인 경우는 긍정문서에 1~4까지의 점수를 부여하였고, 부정문서에는 모두 0을 값을 부여하였다. N인 경우에는 그 반대의 점수를 부여하였으며, PN인 경우는 극성에 관계없이 문서의 질에 따라 1~4까지의 점수만을 부여하였다.

검색 정확도 평가를 위해 테스트를 위한 문서도 학습문서와 같은 방식으로 준비하였다. 테스트 문서는 긍정과 부정문서 중에 각각 100개씩을 선정하였다. 또한 본 실험에서 검색 결과를 비교하기 위한 측정치로서 nDCG(Normalized Discounted Cumulative Gain)를 활용하였다.

표 3. 학습문서의 관련 점수
Table 3. Relevance Score of Learning Documents

polarity / *score*	*P*	*N*	*PN*
4	best 긍정문서	best 부정문서	best
3	good 긍정문서	good 부정문서	good
2	fair 긍정문서	fair 부정문서	fair
1	bad 긍정문서	bad 부정문서	bad
0	부정문서	긍정문서	

2. 실험 결과

표 4는 8개 카테고리로 나눈 각 학습문서 집합에 대해서 본 논문에서 제시한 5가지 특징들의 평균값을 나타낸다. 이 표에서 polarity, length, syntax, speciality, similarity는 차례로 4.1절의 (가)부터 (마)까지 설명한 특징들을 나타낸다. 이 표에서 보는 바와 같이 대부분의 특징들이 문서의 질과 특징 값들이 밀접한 상관관계를 갖는다는 것을 알 수 있다. 하지만 polarity의 경우 긍정문서는 문서의 질과 밀접한 관계가 있는 반면, 부정문서는 상대적으로 관련도가 적은 것을 알 수 있다. 그 이유는 부정적인 표현의 경우 '못하다', '않다'와 같이 부정어를 많이 사용함으로써 unigram의 영향이 상대적으로 작기 때문인 것으로 추정된다. 또한 syntax의 경우도 다른 특징들에 비해 상대적 영향력이 작게 나타났다. 문법적 완

성도를 정확하게 측정하려면 분석 가능한 형태소의 비율뿐만 아니라 형태소들의 문법적 규칙을 정밀하게 측정해야하지만 아직까지 만족할만한 유용한 도구나 방법이 미흡한 것이 사실이다.

표 4. 학습문서의 카테고리별 특징 값 평균
Table 4. Average Feature Scores for Each Category of Learning Documents

features	quality	긍정문서	부정문서
polarity	best	**114**	**- 23**
	good	85	- 11
	fair	66	-7.5
	bad	40	-7
length	best	**82**	**105**
	good	53	77
	fair	38	43
	bad	13	14
syntax	best	**0.96**	**0.97**
	good	0.95	0.95
	fair	0.90	0.90
	bad	0.91	0.90
speciality	best	**3.26**	**2.53**
	good	2.37	1.16
	fair	1.05	0.57
	bad	0.36	0.21
similarity	best	**0.33(0.16)***	**0.11(0.35)**
	good	0.22(0.12)	0.08(0.20)
	fair	0.11(0.08)	0.07(0.15)
	bad	0.09(0.02)	0.01(0.11)

* similarity에서의 값은 best 긍정문서와의 유사도를 나타내며 괄호안의 값은 best 부정문서와의 유사도를 나타낸다.

본 논문에서는 이들의 조합으로 학습한 검색 모델의 정확도를 평가하였다. 표 5는 그 결과를 나타낸다. 이 표에서 *P*와 *N*은 표 3을 근거로 각각 긍정과 부정문서만을 검색한 결과이며, *PN*은 이와 관계없이 문서의 질에 따른 검색 결과이다. 성능평가 결과를 보면 *P*의 경우 polarity+length+speciality와 같이 3개의 특징들로 학습한 경우가 가장 좋은 검색 정확도를 보였으며, *N*의 경우는 polarity+length+speciality+similarity를 사용한 경우가 가장 좋은 정확도를 보였다. *PN*의 경우는 polarity+length+similarity가 가장 좋은 검색 정확도를 보였다. 이 실험에서 본 바와 같이 *P*, *N*, *PN* 모두의 경우 공통적으로 polarity와 length가 포함된 학습 결과가 가장 좋은 성능을 보였다. 검색의 목표가 오피니언 문서라는 점을 감안한다면 polarity가 중요한 역할을 한다는 것은 쉽게 이해할 수 있다. length가 중요한 특징으로 평가된 것에 대해서는 학습 문서가 단문이라는 특징에 기인한다. 일반적인 경우 문서의 길이가 문서의 질을 평가하는데 큰 영향이 없다. 하지만 네이버 영화평이나 트위터와 같이 단문만을 허용하는 경우, 지나차게 짧은 문서들이 많을 수 있어 상대적으로 긴 문서일수록 좋은 문서일 가능성이 높아지게 된 것으로 해석할 수 있다.

표 5. 특징 조합에 따른 rankSVM의 검색 성능 비교
Table 5. Search Result Comparison of rankSVM

features	NDCG@10		
	P	*N*	*PN*
polarity+length+specialty	**0.840**	0.744	0.888
polarity+length+syntax	0.759	0.702	0.744
polarity+length+similarity	0.823	0.780	**0.903**
syntax+speciality+similarity	0.527	0.664	0.626
polarity+length+speciality+similarity	0.838	**0.812**	0.839
All features	0.838	0.783	0.839

syntax를 포함하여 학습한 경우는 상대적으로 큰 영향이 없는 것으로 나타났다. 본 논문에서 수집한 학습문서의 경우 수작업으로 분류할 때 fair나 bad로 분류된 문서에 대해서 문법적으로 오류가 많은 경우를 흔히 발견할 수 있었다. 하지만 본 논문에서 측정한 정량적 수치는 단지 형태소분석기의 미분류 비율만을 측정하였으며, 이 값만으로는 문법적 완성도에 대한 충분한 측정이 불가능했던 것으로 판단된다.

본 논문에서 실험한 문서들은 모두 오피니언 문서라는 가정에 출발하였다. 그러나 오피니언 문서와 객관적 문서가 혼재된 경우에는 그 결과가 얼마든지 달라질 수 있다. 예를 들어 트위터와 같이 오피니언 문서와 객관적 문서가 동시에 존재하는 상황에서 오피니언 문서를 검색할 경우에는 이 방법을 그대로 적용하기에는 한계가 있다. 그 이유는 모든 문서들 중에서 객관적 문서들을 배제하고 오피니언 문서만을 선별하는 과정이 필요하기 때문이다. 이를 위해서는 전처리 과정으로 오피니언 문서와 객관적 문서를 분류하거나, 검색 모델을 생성하는 기계

학습 과정에서 오피니언 문서의 선별을 위한 특징들을 추가적으로 개발하는 과정이 요구된다.

VI. 결론

본 논문에서는 SNS와 같은 불특정 다수에 의해 게시된 단문 오피니언 문서를 효율적으로 검색하기 위한 방법을 제안하였다. 제안된 방법은 감성 극성을 분류하는 방법을 검색을 위한 하나의 특징으로 취급하고, 문서의 질을 평가할 수 있는 여러 가지 요소들을 통합하여 감성 분류와 검색을 하나의 프레임워크 내에서 해결하도록 설계하였다. 또한 기계학습 랭킹 기법을 사용하여 검색 결과의 객관성을 보장하였으며, 실험을 통하여 검색 종류에 따라 최적의 특징 조합을 탐색하였다.

본 논문에서 제안한 방법의 가장 큰 한계는 객관적인 성능을 검증하는 실험 데이터가 없다는데 있다. 향후에는 이러한 한계를 극복하기 위한 실험 데이터의 준비가 필요할 것으로 판단된다. 아울러 본 논문이 결과를 바탕으로 트위터와 같은 보편적인 SNS에서의 오피니언 검색에 관한 적용방안도 연구할 계획이다.

References

[1] H. Kim and J. Chang, “Improving Naive Bayes Text Classifiers with Incremental Feature Weighting”, Journal of Korea Information Processing Society, Vol. No. 5, pp.457-464, 2008.

[2] J. Chang and I. Kim, “An Experimental Evaluation of Short Opinion Document Classification Using A Word Pattern Frequency”, Journal of the Institute of Internet, Broadcas ting and Communication, Vol. 12, No. 5, 2012.

[3] J. Kim, S. Lee, and H. Yong, “Automatic Classification Scheme of Opinions Written in Korean ”, Journal of KIISE: Database, Vol. 38, No. 6, 2011.

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

[5] http://lucene.apache.org/nutch/

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

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

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

[9] B. Liu , M. Hu, and J. Cheng, “Opinion observer: analyzing and comparing opinions on the Web”, Proceedings of the 14th international conference on WWW, pp. 10-14, 2005.

[10] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, “Red Opal: Product-Feature Scoring from Reviews”, Proceedings of the 8th ACM conference on Electronic commerce, pp. 11-15, 2007.

[11] Xiaowen Ding and Bing Lui, “The Utility of Linguistic Rules in Opinion Mining", Proceedings of SIGIR 2007, pp. 811-812, 2007.

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

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

[14] T. Liu, Learning to Rank for Information Retrieval, now Publisher Inc. 2009.

[15] T. Joachims, "Optimizing Search Engines using Clickthrough Data", Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, 2003

[16] H. Yu, Y. Kim, and S. Hwang, “RV-SVM: An Efficient Method for Learning Ranking SVM”, Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, Volume 5476, pp 426-438, 2009.

[17] http://en.wikipedia.org/wiki/PageRank

[18] X. Huang and W. B. Crott, “A Unified Relevance Model for Opinion Retrieval”, Proceedings of CIKM ‘09, 2009.

[19] B. Li, L. Zhou, Shi Feng, and K. Wong, “A Unified Graph Model for Sentence-based Opinion Retrieval”, Proceedings of 48th Annual Meeting of the Association for Computational Linguistics, pp. 1367-1375, 2010.

[20] W. Zhang, C. Yu, and W. Meng, “Opinion Retrieval from Blogs”, Proceedings of CIKM ‘07, 2007.

[21] Y. Yang and J. Pedersen, ″A comparative study on feature selection in text categorization,″ Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.

[22] C. Park, D. Seong, K. Lee, “Automatic IPC Classification for Patent Documents using Machine Learning″, Journal of Korean Institute of Information Technology, Vol. 10, No. 4, 2011.

[23] 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.

저자 소개

장 재 영(정회원)

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

<관심분야 : 데이터베이스, 정보검색, 데이터마이닝>

※ 이 논문은 2011년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임. (과제번호: NRF-2011-0022445).