단어-평점 상관관계 분석을 통한 단문 영화평 평점 예측

Prediction of Rating Score from Short Comments on Movies using Word-Rating Correlation Analysis

윤두밈, Dumim Yoon*, 김경중, KyungJoong Kim**

요약 최근, 인터넷 상의 각종 의견정보를 분석하는 opinion mining 연구가 활발히 진행 중이다. 온라인 쇼핑몰, Social Network 서비스 등을 통해 사용자의 의견 정보는 폭발적으로 증가하고 있고, 이를 효과적으로 분석하기 위한 기술개발이 필요한 시점이다. 하지만, 영문을 중심으로 opinion mining 연구가 이루어지면서, 한국어를 대상으로 하는 기술 개발이 부족한 상황이다. 본 연구에선 40자의 짧은 영화평을 대상으로 사용자가 매긴 평점을 예측하는 기술을 제안한다. 학습 데이터로부터 단어와 평점사이의 관계를 분석하고, 이를 이용하여 새로운 영화평에 대한 평점을 자동으로 예측하는 것이다. 이러한 기술은 사용자가 일일이 영화평에 점수를 매기지 않더라도 높은 수준의 정확도로 예측을 가능하도록 한다. 네이버에서 제공한 영화 박쥐에 대한 10000건의 영화평과 평점 정보를 토대로 제안한 방법의 정확도를 평가해 보며, 수작업으로 단어선정을 한 경우와 형태소 분석기를 이용하여 대상단어를 자동으로 추출한 경우를 비교 평가한다.

Abstract Recently, opinion mining researches are making progress with various opinion information analysis on internet. User's opinion information are explosively increasing through online shopping malls and social network services. So, we need technology for effective analysis. However, most opinion mining researches have focused on English, and Korean's one is suffered from the lack of technology development. This paper proposes a rating prediction method based on 40-characters Korean short movie reviews. It is able to automatically predict new movie reviews' rating using word-rating correlation analysis. This method enables to predict the user's rating accurately wihtout user's specific rating report. We tested 10000 reviews of movie 'thirst' from Naver by comparing two dictionaries designed manually and automatically by Korean Morphological Analyzer.

핵심어: Opinion mining, Short comment, Movie reviews, Korean Morphological Analyzer

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(2010-0012876) 및 뇌과학 원천기술개발사업임(2010-0018948).

*주저자 : 세종대학교 컴퓨터공학과 교수 e-mail: krad@hanmir.com

**공동저자 : 세종대학교 컴퓨터공학과 교수 e-mail: kimkj@sejong.ac.kr

1. 서론

인터넷은 폭발적으로 증가하는 사용자 의견으로 넘쳐나고 있다. 기존의 객관적인 정보와 달리 사용자 의견은 주관적인 성향을 가지고 있으며, 이를 자동으로 분석하려는 연구도 폭넓게 이루어져 왔다[1]. 즉, 사용자의 의견이 긍정적인 것인지, 부정적인 것인지를 자동으로 분석하여 사용자가 의견을 굳이 읽어보지 않더라도 시스템이 자동으로 분류해 줄 수 있는 기능을 제공하려 하였다.

본 논문에서는 40자로 제한되어 있는 단문의 영화평을 대상으로 사용자의 평점을 예측하는 기술을 제안한다. 데이터는 Naver에서 공개한 오피니언 마이닝 데이터를 이용하였다[4]. 본 데이터는 영화 '박쥐' 와 '해운대' 에 대해 각각 10000건의 40자 이내의 영화평과 1~10점 사이의 평점 정보를 제공하고 있다. 본 논문에서는 사용자의 영화평으로부터 평점을 예측하기 위한 단어사전을 자동으로 구축하는 방법을 제안한다.

3. 단문영화평 기반 평점 예측

상품평등의 일반적인 Opinion Mining에 있어서는 의미사전을 이용한 방식[2]이나 평가의 특징을 사용하는 방식[3]으로 주로 서술형식의 문장에서 어떻게 적절한 특징을 추출해 요약하는가에 중점을 두는데 반해, 영화의 한줄 평은 길이가 짧고 함축적이며 은유적인 표현과 더불어 해당 영화 제목, 배우, 감독, 동원 관객수, 동시 상영 영화 등, 직접적으로 주어지지 않는 배경 정보가 많이 포함되어 있다.

3.1 전처리

Naver에서 제공하는 데이터는 XML형태로 되어 있으며, 각 건별 영화 평점 (1~10)과 40자이내의 영화평이 제공되고 있다. "박쥐" 와 "해운대" 두편의 영화에 대해 각각 10,000건의 데이터가 제공되고 있다. 본 논문에서는 전처리를 통해 영화평과 평점 정보를 추출하도록 한다.

3.2 중요단어의 추출

한국어는 영어와 달리 단어와 조사가 결합되어 있는 형태이기 때문에, 분석이 더욱 어렵다. 일반적으로 전체 단어 중에서 일부만 분석에 유용한 단어이기 때문에, 핵심적인 단어를 추출하여 평점과의 관계를 분석하는 것이 필요하다. 본 연구에선 두 가지 방법을 사용하여 이러한 과정을 수행하였다. 첫 번째 방법에선 시스템 개발자가 영화평을 수작업을 통해 분석하고, 핵심적인 단어 200~300개 정도를 추출하였다. 이후, 영화평과의 상관관계 분석을 통해 해당 단어의 극성을 분석하였다. 두 번째 방법은 형태소 분석기를 이용하여 영화평을 모두 품사별로 구분하고, 이 중에서 명사와 동사를 추출한 이후에 이를 대상으로 단어-평점 관계를 분석하는 것이다.

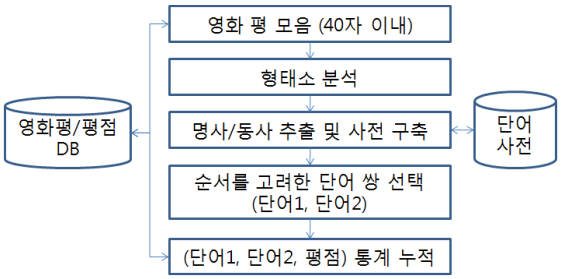

그림 1 단어-평점 상관관계 분석 과정

3.2 단어-평점 상관관계 분석

본 논문에서는 두 단어 쌍과 평점사이의 관계를 분석한다. 하나의 단어 쌍을 (w1, w2)로 표기하면, 영화 평에서 w1은 항상 w2보다 먼저 나와야 한다. 예를 들어, "지금까지 본 영화 중 최고!" 라는 영화평이 있으면, (지금, 영화), (영화, 최고) 등이 하나의 단어 쌍을 이룬다. 각각의 단어 쌍에 대해 전체 영화평을 분석하여 매치하는 경우마다 평점을 합산한 후, 최종적으로 평균 평점을 계산한다. 즉, (지금, 영화)와 일치하는 영화평이 총 10개가 있고, 그들의 평점 평균이 4점이라면, (지금, 영화)의 평균평점은 4가 된다. 평점의 표준편차도 중요한 요소가 될 수 있기 때문에 함께 고려하도록 한다.

$$Avg(w1,w2) = \frac{\sum_{i=1}^{N} Rating_i}{N} \quad (1)$$

$$Std(w1,w2) = \frac{1}{N}\sqrt{((Rating_i) - Avg(w1,w2))^2} \quad (2)$$

위 수식은 각 단어 쌍에 대해 평균 평점과 표준편차를 구한 것을 보여준다. 이 수식에서 N은 전체 영화평 중에서 (w1, w2) 쌍을 포함하는 영화평의 개수를 의미한다. 이러한 통계 분석을 통해 해당 단어 조합의 평균적인 평점을 알 수 있으며, 얼마나 신뢰할 수 있는지 표준편차를 통해 구할 수 있다. 이러한 통계 데이터를 기반으로 새로운 영화평에 대한 평점을 예측하도록 한다.

표 3. 단어 쌍의 예와 평균 평점 및 표준편차

	Avg(w1,w2)	Std(w1,w2)
(지금, 영화)	4	5
(영화, 최고)	10	2
(최고, !)	8	1

3.4 평점 예측

평점은 영화평에 포함되어 있는 모든 단어 쌍에 대해 이

미 구축한 Avg, Std 데이터를 활용하여 수식으로 계산한다.

3. 실험 및 결과

본 논문에서는 '박쥐 " 영화에 대한 영화평을 사용하였다. 그림 2에 나오는 것처럼 영화 평점은 대부분 1점과 10점으로 양분되어 있는 것을 확인할 수 있다. 총 10000개의 영화평 중에서 65%가 1점 또는 10점이었다. 본 논문에서는 이러한 특성을 토대로 1점과 10점 데이터만을 사용하였다.

영화평 데이터로부터 영화평점을 예측할 때 사용할 단어를 추출하는 것이 필요하다. 첫 번째 방법은 수작업을 통해 중요한 단어를 선택하였다. 이 경우, 중요한 단어를 확인하는 과정이 수작업을 통해 이루어지기 때문에, 사전 구축에 많은 시간이 소요하고, 선택할 수 있는 단어의 수도 제한이 따른다. 하지만, 일반적인 형태소 분석으로 추출 할 수 없는 단어나 문법을 지키지 않아 정확히 분별하기 어려운 단어 등을 찾아낼 수 있는 장점이 있다. 두 번째 방법은 형태소 분석기를 이용하여 명사와 동사만을 골라낸 후 이들을 이용하는 방법이다[5].

본 논문에서 제안한 방법으로 영화평을 예측할 경우, 분석이 불가능한 한줄 평이 발생할 수 있다. 예를 들어, 한줄평에 사용한 단어가 사전에 없을 경우이다. 수작업을 하는 경우, 전체 단어 중에서 일부분만 포함하기 때문에 이러한 경우가 발생할 수 있으며, 형태소 분석을 하는 경우에도, 형태소 분석이 실패하여 단어를 추출하지 못한 경우도 있을 수 있다.

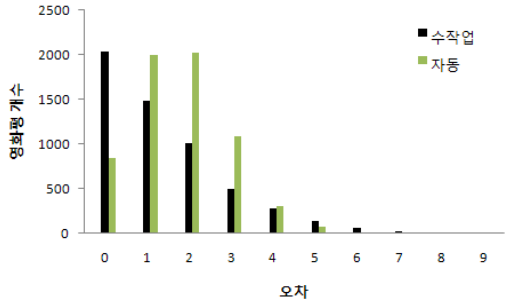

그림 2 '박쥐' 데이터에 대한 평점 예측 오차 분포도

그림 3에서 볼 수 있듯이, 수작업을 한 경우와 자동화 한 경우의 오차 분포에 큰 차이가 있다. 수작업을 한 경우 오차가 0인 경우가 상대적으로 많았으나, 자동화한 경우 오차가 1~3인 경우가 많았다. 수작업을 한 경우 누락 샘플의 수가 967개로 전체 샘플 6481개 중에서 15%정도였고, 자동화 경우 누락 샘플의 비율은 2% 정도로 매우 낮았다. 평균 오차의 경우 수작업을 한 경우 1.31이었으며, 자동화 한 경우 1.73이었다. 수작업을 한 경우가 비교적 낮게 나왔으나, 높은 누락 비율을 보이는 문제가 있었다. 자동화 한 경우에도 평균 오차 2미만으로 신뢰도 높은 예측을 할 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 짧은 영화평정보를 토대로 사용자의 평점을 예측하는 연구를 수행해 보았다. 가장 중요한 부분인 단어 추출 부분을 수작업을 통해 하는 경우와 자동화 하는 경우를 상호 비교하였으며, 단어 쌍의 평균평점과 표준편차 데이터를 이용하여 평점을 예측하는 방법을 제안하였다. 제안한 방법을 통해 전체 과정을 자동화 할 수 있는 가능성을 확인하였으며, 2미만의 오류 범위 이내에서 평점을 예측할 수 있음을 확인하였다.

향후 연구 과제로는 현재 사용된 형태소 분석기를 이용한 자동화의 경우 명사와 동사만을 이용하였는데, 명사나 동사 혹은 이번 실험에 포함되지 않은 조사들 중 어느 품사가 감성 표현에 더 중요한 영향을 가지고 있는지를 조사할 예정이다. 또한, 본 실험의 결과를 확장하여, '박쥐' 로부터 얻은 단어-평점 관계를 이용하여 '해운대' 의 평점을 예측하는 등의 일반화 성능 테스트가 필요하다.

참고문헌

[1] B. Pang, and L. Lee, "Opnion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[2] 명재석, 이동주, 이상구, "반자동으로 구축된 의미사전을 이용한 한국어 상품평 분석 시스템," 정보과학회 논문지: 소프트웨어 및 응용, vol. 35, no. 6, pp. 392-403, 2008.

[3] 장재영, "온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘," 한국전자거래학회지, vol. 14, no. 4, pp. 19-33, 2009.

[4] Naver Opinion Mining Data, http://lab.naver.com

[5] 강승식, 한국어 형태소 분석과 정보검색, 홍릉과학출판사, 2002.