소셜데이터 감성분석을 통한 사용자의 호감도 분석

이민규* · 손효정** · 성백민*** · 김종배°

*숭실대학교 소프트웨어특성화대학원

Favorable analysis of users through the social data analysis based on sentimental analysis

Min-gyu Lee* · Hyo-jung Sohn** · Baek-min Seong*** · Jong-bae Kim°

*Graduated Soongsil university

E-mail : marse101@naver.com* hyojung.sohn@gmail.com** feeling127@naver.com*** kjb123@ssu.ac.kr°

요 약

최근 폭발적으로 증가하는 SNS서비스의 상업적으로 이용하려는 움직임이 활발하다. 따라서 본 논문은 실시간 SNS 환경에서 제조기업과 제품의 평판에 관련된 정보를 정확하게 분석 할 수 있는 방안을 제시한다. 크롤링 방식으로 수집된 SNS의 텍스트 데이터들에 대한 형태소 분석을 수행하여 단어 간 연관성을 파악한다. 또, 문장에서 추출된 형태소는 구축된 감성사전을 통해 통계적으로 분석하여 이를 시각화 하여 보여준다. 이때, 추출된 단어가 감성사전에 존재하지 않을 경우 이를 자동으로 추가하는 기법을 제안한다.

ABSTRACT

Recently it is used commercially to actively move the data from the SNS service. Therefore, we propose a method that can accurately analyze the information related to the reputation of companies and products in real time SNS environment in this paper.Identify the relationship between words by performing morphological analysis on the text data gathered by crawling the SNS scheme. In addition, it shows the visualization to analyze statistically through a established emotional dictionary morphemes are extracted from the sentence. Here, if the extracted word is not exist in sentimental dictionary. Also, we propose the algorithm that add the word to emotional dictionary automatically.

키워드

감성분석, SNS, 소셜, 소셜데이터, 감성단어, 감성사전

Ⅰ. 서 론

최근 폭발적으로 증가하는 SNS서비스의 상업적으로 이용하려는 움직임이 활발하다. 한가지 예로, 코카콜라의 'SNS 데이터 활용을 통한 가치향상 노력'은 이용자들의 호불호가 빠르게 반영되는 SNS 데이터를 실시간으로 적용하여 고객 맞춤형 서비스를 제공하였다. 코카콜라는 SNS를 이용한 마케팅으로 소비자의 반응을 빠르고 쉽게 파악하여 기업이 개선해야 할 점이 무엇인지 쉽게 알 수 있는 환경으로 변화하였다. 그로 인해 소통하는 브랜드라는 이미지로 탈바꿈 하여 브랜드가치를 향상시킬 수 있었다. [1]

본 논문은 실시간 SNS 환경에서 제조기업과 제품의 평판에 관련된 정보를 정확하게 분석 할 수 있는 방안을 제시한다. 이를 위해 크롤링 방식으로 수집된 SNS의 텍스트 데이터들에 대한 형태소 분석을 수행하여 단어 간 연관성을 파악한다. 크롤링은 구글에서 제공하고 자바로 작성된 오픈소스 Crawler4j[2]을 사용해서, 트위터, 페이스북 등 SNS 웹페이지에서 해당 제품에 대한 글들을 수집한다.

다음으로, 문장에서 추출된 형태소는 사전에

구축된 감성사전을 통해 김정호등[3]이 제안한 방법을 이용하여 통계적으로 분석한다.

이때, 추출된 단어가 감성사전에 존재하지 않을 경우 국어사전에서 단어의 의미를 추출하고 형태소 분석해서 의미를 분석하여 자동으로 추가하는 기법을 제안한다.

최종적으로 분석된 감성수치는 그래프를 이용하여 시각화한다.

II. 관련 연구

형태소 분석 방법에 관한 연구로는 서울대학교의 지능데이터시스템 연구실의 꼬꼬마프로젝트가 있다. 꼬꼬마 형태소 분석기는 홈페이지에서 형태소 분석기 라이브러리와 사전 데이터를 배포하고 저작권은 GPL 2.0을 따른다.[4] 본 연구에서는 위의 꼬꼬마 형태소 분석기를 수집된 텍스트에 대한 형태소 분석 도구로 활용한다.

감성분석 기법에 관한 연구로는 Zhongchao가 제시한 감성분류 방법과 김정호등이 제시한 한국어 특성을 고려한 감성 분류 방법 등이 연구되고 있다. Zhongchao가 제시한 방법은 영어 문법에 맞게 작성되어 있으나 한국어의 복잡한 성격에는 그대로 적용하기 힘들다. 따라서 본 연구에서는 Zhongchao가 제시한 방법을 인용하여 한국어 특성에 맞게 수정된 김정호의 방법을 사용한다.

시각화 방법에 관한 연구는 조은희[5]의 실시간 수치데이터의 이미지 기반 시각화 방식에 대한 연구가 있고 본 연구에서는 단순히 감성 수치를 점수로 보여주고, 이를 그래프로 표출하는 방법을 쓴다.

III. 감성분석을 통한 사용자의 호감도 분석

3.1 시스템 구성도

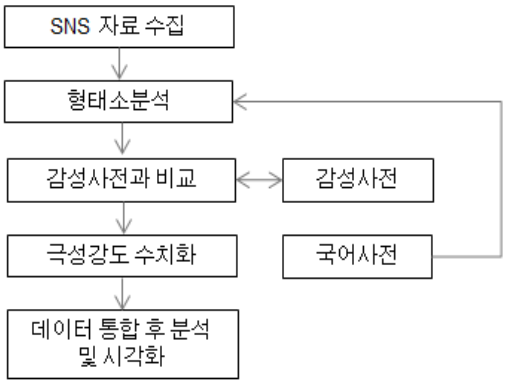

그림 1. 시스템 구성도

전체적인 호감도를 분석하는 절차는 그림1과 같다. 트위터, 페이스북등 SNS 사이트에서 사용자의 텍스트데이터를 크롤링하여 형태소 분석을 한다. 미리 저장해 놓은 감성사전의 단어들과 형태소분석을 통해 도출된 형태소를 김정호등이 제시한 방법을 이용해 극성 강도를 수치화 하여 사용자의 긍정 또는 부정적 의견을 판별한다. 하지만, 만약 형태소분석을 통해 도출된 형태소가 미리 저장해놓은 감성사전의 단어와 일치하는 것이 없다면 제안하는 알고리즘을 통해 감성사전에 추가한다.

3.2 문장의 극성 강도

극성 강도는 개발자가 직접 정할수도 있지만 이는 극히 주관적이기 때문에 좋은 방법이라 할 수 없다.

$$w_i \begin{cases} \log\left(\frac{T_{ai}}{T_{di}}\right) & T_{ai} \neq 0 \text{ and } T_{di} \neq 0 \\ C+\log\left(\frac{T_{ai}+1}{T_{di}+1}\right) & T_{ai} = 0 \text{ or } T_{di} = 0 \end{cases}$$

따라서 이러한 주관적인 문제를 해결하기위해 김정호등은 Zhongchao[6]가 제안한 방법을 인용했다. 위 식에서 T_{a_i}는 긍정문장에 나타난 긍정적 단어 빈도수 이고 T_{d_i}는 부정문장에 나타난 부정적 단어 빈도수 이다. w_i는 감성강도를 나타내며 w_i>0 이면 긍정의 감성을 가지고 w_i<0이라면 부정의 감성을 가지고 있는 것으로 판별한다.

3.3 감성사전

감성사전은 미리 지정하는 감성단어의 집합이고 이 사전은 박인조[7]등이 제안한 한국어 감정단어 목록을 이용하여 작성 한다. 그는 연세대 정보개발연구원에서 제작한 '현대 한국어의 어휘빈도'로부터 감성단어를 추출 하였으며, 감성단어의 선정에는 감성연구자 10명이 참가하였으며 434개의 감정단어들로 구성된 목록을 완성하였다. 위의 연구에서 긍정 또는 부정적 단어의 목록을 긍정단어 감성사전과 부정단어 감성사전을 작성한다.

3.4 감성사전 추가 알고리즘

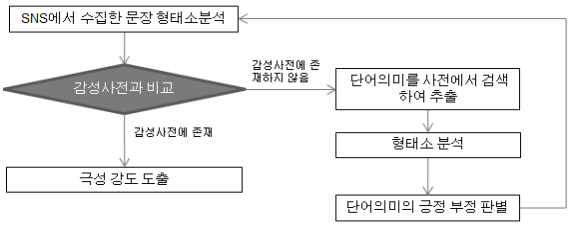

그림 2. 감성사전 추가 알고리즘

먼저 그림1에서 SNS 자료수집 후 형태소분석을 하고 그림 2와같이 감성사전 추가 알고리즘에 의해 감성사전과 비교를 할 때 만약 감성사전에 형태소가 존재하지 않으면, 존재하지 않는 단어의 의미를 국어사전에서 추출하여 형태소분석을 하고 이를 다시 감성사전과 비교, 수치화 하여 긍정의 단어인지 부정의 단어인지 판별 후 감성사전에 추가한다.

3.5 패턴의 구성

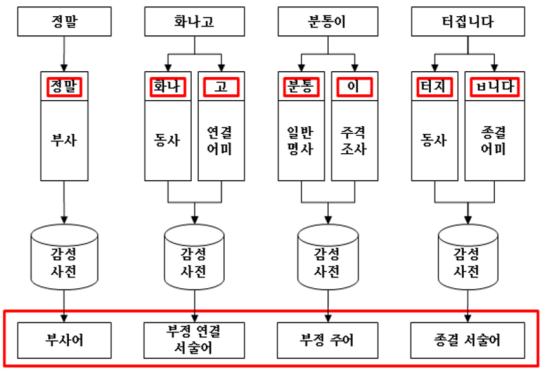

그림 3. 김정호등이 사용한 문장성분으로 구성된 구문패턴 추출예제

Zhongchao가 제시한 방법은 영어 문법에 관해 감성분석 방안을 제시하였기에 그대로 사용하기에는 무리가 있다. 한국어는 하나의 어절이 여러 형태소들로 이루어져 있기 때문에 어절을 이루는 모든 형태소들의 품사를 패턴에 이용하면, 생성될 수 있는 패턴 조합의 수가 너무 많아져 계산 양이 증가하고 정확률이 떨어 질 수 있다는 것이다. 그래서 김정호등은 문장성분으로 구성된 구문 패턴 방법을 제안 하였다. 그림 3은 패턴을 추출하는 예를 나타낸다. 문장에 해서 형태소 분석 및 감성사전 검색을 통해 극성을 나타내는 문장성분 패턴을 추출한다. 구체적으로 설명하면, 부사는 부사어, 연결어미는 연결 서술어, 주격조사는 주어, 종결어미는 종결 서술어를 나타내고, 동사 '화나-'와 명사 '분통'은 감성사전 에서 극성을 나타내는 단어이기 때문에 이 단어를 포함하는 문장성분에 극성을 표시하여 각각 '부정 연결 서술어', '부정 주어'로 나타내었다.

Ⅳ. 결론

본 논문에서는 실시간 SNS 환경에서 제조기업과 제품의 평판에 관련된 정보를 정확하게 분석 할 수 있는 방안을 제시하였다. 크롤링 방식으로 수집된 SNS의 텍스트 데이터들에 대한 형태소 분석을 수행하여 단어 간 연관성을 파악하였다. 또, 문장에서 추출된 형태소는 구축된 감성사전을 통해 통계적으로 분석하여 이를 시각화 하여 보여주었다. 이때, 추출된 단어가 감성사전에 존재하지 않을 경우 이를 자동으로 추가하는 기법을 제안하였다. 본 연구의 결과는 제품사용자의 호감도를 더욱 쉽고 정확하게 판별 할 수 있도록 하기 위한 방법으로 활용될 수 있다. 위와 같은 자동화 알고리즘을 활용함으로써 기업들은 소비자의 반응에 더욱 빠르게 대응 할 수 있어서 소비자 요구에 맞춰 가치창출이란 효과를 얻을 수 있다. 그러나, 본 연구는 아직 실험단계에 그치고 있어, 향후 문장패턴 구성 방식에 대한 연구가 추가로 필요하고 감성추출의 성능테스트가 필요하다.

참고문헌

[1] http://blog.m-adfactory.com/220133253111
[2] https://code.google.com/p/crawler4j/
[3] 김정호, 김명규, 차명훈, 인주호, 채수환, 한국어 특성을 고려한 감성 분류,감성과학, vol.13, no.3, 339-458, 2010
[4] http://kkma.snu.ac.kr/
[5] 조은희,김현욱 , 실시간 수치데이터의 이미지 기반 시각화 방식에 대한 연구, 2011
[6] Zhongchao Fei, Jian Liu, Gengfeng Wu, Sentiment Classification Using Phrase Patterns, In The Fourth International Conference on Computer and Information Technology (CIT' 04), 1147-1152, 2004.
[7] 박인조, 민경환, 한국 감정단어의 목록 작성과 차원 탐색, 성격 및 사회심리학지, vol.19, 109-129, 2005