

키워드 기반 분산 SNS 검색 및 오피니언 마이닝 시스템

윤한중*, 석상기*

*서울과학기술대학교 컴퓨터공학과
e-mail : {christ26y, sksuk}@seoultech.ac.kr

Distributed SNS Crawling and Opinion Mining System

Han-Jung Youn*, Sang-Kee Suk*

*Dept. of Computer Science & Engineering, Seoul National University of Science and Technology

요약

제안된 시스템은 다양한 소셜 네트워크에서 사용자가 입력한 키워드를 기반으로 데이터를 수집하여 형태소 분석을 거쳐 사용자에게 통계정보 및 키워드에 대한 오피니언 마이닝 결과를 제공한다. SNS 상에 수많은 정보들이 저장되는데, 이를 이용하는 과정에서 단편적인 정보밖에 얻을 수 없는 비전문적인 사용자에게 유용한 데이터를 제공하기 위해 Opinion Mining 및 다양한 통계적 분석을 통해 키워드에 대한 시각화 정보를 출력한다.

1. 서론

최근 소셜 네트워크 환경(SNS)은 지속적으로 성장하며 영역을 넓히고 있다. 온라인상에서 중요한 주제를 가지고 토론하거나, 서로의 취미를 공유하는 등, 이미 SNS는 우리의 일상에 깊게 파고들었다. 이 과정에서 다양한 종류의 SNS가 개발되어 각자의 특징에 따라 사용자들은 서비스를 선택하여 사용하고 있다. 하나의 플랫폼에서만 데이터를 수집하는 것에 비해 다양한 환경에서 정보를 수집할 수 있다면, 이는 같은 키워드에 대해 더 넓은 스펙트럼의 데이터를 가질 수 있음을 의미한다.

본 논문에서는 키워드 기반의 분산 소셜 네트워크 서비스 크롤링을 통해 데이터를 수집하고 정리하여 감성 분석에 활용하거나, 사용자에게 워드 클라우드 및 데이터의 시각화를 통해 유의미한 정보를 제공하는 도구를 제안한다 [1].

2. 관련연구

2.1 SNS 데이터 수집

소셜 네트워크의 데이터 수집은 이미 많은 연구에서 시도하고 있다. 트위터나 페이스북을 통한 텍스트 데이터 및 인스타그램, 플리커 등을 활용한 이미지 데이터까지 이슈 분석 및 감성분석에 이용된다. 본 연구는 각각의 SNS에서 다양한 유형의 데이터를 수집하여 시스템의 사용자에게 복합적인 시각데이터와 통계 정보를 제공하고자 하였다 [2, 3].

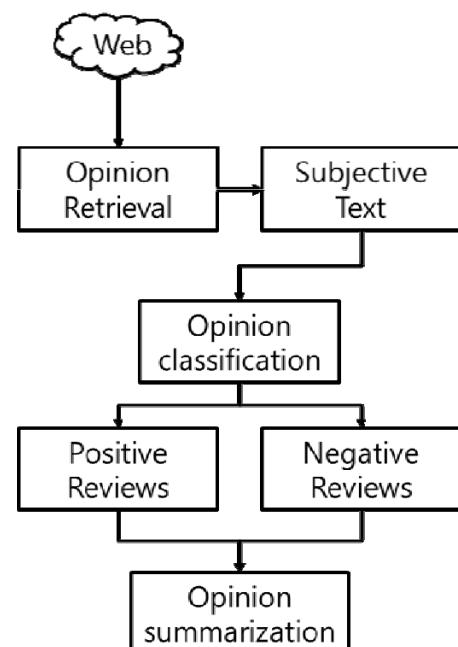
2.2 자연어 처리

본 논문은 한국어 기반으로 연구가 진행되며, 이 과정에서 보편적으로 자연어 처리에 사용되는 NLTK를 이용하지만, NLTK는 한국어 지원이 미흡하여 한국어 형태소 분석을 위하여 KoNLPy 라이브러리를 추가로

활용하여 한글 문장을 분석, 계산한다 [4].

2.3 오피니언 마이닝

감성 분석으로 지칭되기도 하는 오피니언 마이닝은 키워드에 대한 사용자의 반응과 의견을 분석하는 기술이다. 대상의 의견이 긍정적, 부정적, 또는 중립적인지 파악하며 또한 문장에 사용된 형용사의 강도에 따라 감정의 강도를 분석한다. 오피니언 마이닝은 하나의 문서, 문장, 단어 단위에서 다양하게 실행될 수 있으며, 그림 1에서 보이는 것과 같이 의견 수집, 의견 분류, 의견 요약의 세 단계로 구성되어 있다 [5].

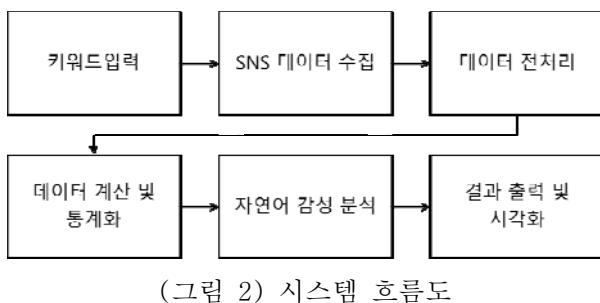


(그림 1) 오피니언 마이닝의 구조

3. 본론

3.1 시스템 소개

그림 2는 제안한 시스템의 작동방식과 주요 진행 과정을 도식화한 흐름도이다. 그림 2에서 보인 것과 같이 본 연구는 비전문적인 사용자로부터 키워드를 입력 받아 그것을 기반으로 데이터를 다양한 소셜 네트워크 서비스에서 수집하고, 수집된 데이터를 계산하고 분석하기 쉽게 처리한 뒤 이를 이용해 연관 키워드를 도출하고, 문장 주변의 단어를 감성어 분석을 통해 키워드에 대한 긍정 또는 부정적인 정도를 파악한다. 이 후 그래프, 도표, 워드 클라우드 등을 통해 계산된 결과를 시각화하는 시스템을 제안하였다.



3.2 시스템 구현

본 논문은 복수의 SNS로부터 데이터를 수집하여 키워드 기반으로 정리, 분석하는 프로그램을 제안하였으며, 사용자가 검색하고자 하는 키워드를 입력하면 사용자에게 유용하고 신뢰성이 있는 연관 키워드 목록과 키워드에 대한 사람들의 반응을 파악할 수 있도록 하는 연구이다.

<표 1> 추출된 키워드 결과물

트위터 원문	추출 키워드
애플 맥 앱스토어 장애 발생... 보안 인증서 갱신으로 인해 앱 재설치 필요할 수도	애플 맥 스토어 장애 발생 보안 인증서 갱신 재 설치 필요
[현장] 삼성 갤럭시 S7·S7 엣지 출시...DSLR급 화질에 방수까지	현장 삼성 갤럭시 S7 출시 금 화질 방수
핸드폰이 자꾸 꺼져서 서비스 센터를 찾아갔다. 추우면 베테리가 남아있어도 방전이라고	핸드폰 꺼져 서비스 센터 찾아가 추우 베테리 방전

현재 연구는 트위터의 데이터를 수집하여 자연어 처리를 거쳐 분석된 통계 데이터를 문자열로 출력하는 과정에 있으며, 이 과정에서 Python 2.3과 SQLite를 이용하여 데이터를 수집하고 저장한다. 문자 데이터는 NLTK 및 KoNLPy 라이브러리를 이용하여 형태소 분석 및 TF*IDF를 이용해 단어의 중요도와 빈도를 계산하여 단어 벡터를 생성한다. 단어벡터는 감성어 분석 알고리즘을 이용하여 각 단어 별로 가중치를 부여하여 트윗의 긍정/부정을 결정한다 [6, 7]. 표 1은 삼성과 애플을 키워드로 검색한 결과 중 일부를 첨부한 것으로, 외래어, 문법, 합성어 등의 문제로 단어가 깨지거나 의미 불명의 형태소로 추출되는 문제가 확인되었다. 향후 형태소 분석 과정에서 일부

키워드를 직접 입력하거나 형태소 분석 알고리즘을 일부 수정하는 등의 추가 변화가 필요하다.

4. 결론 및 한계

본 연구의 결과물은 다양한 분야에서 활용 가능한 크롤러 및 감성분석 시스템으로, 소셜 네트워크를 기준으로 작동하기 때문에 현대 사회에서 SNS 플랫폼이 지속적으로 활용되기 때문에, 제품이미지 분석, 사회 이슈에 대한 분석 등을 위해 사용 가능한 시스템이다. 그러나, 제안된 시스템은 사용자의 키워드가 입력된 뒤에 SNS 크롤링을 시작하기 때문에, 만족스러운 계산 속도를 보장하지 못하며, SNS마다 게시글의 형태가 다르고, 이미지 처리에 대한 계산이 배제되어 있다. 이로 인해 사진 중심의 SNS는 본 모델에서는 검색 대상으로 포함하기 어려운 문제가 있다.

향후 연구에서는 지속적으로 무작위 데이터를 수집하여 DB에 저장하고, 사용자가 키워드를 입력하는 경우 키워드에 대해 DB 상에서 검색을 진행하여 시간의 흐름에 따른 키워드의 빈도 등 추가적인 정보를 수집할 수 있을 것으로 보인다. 또한, 수집된 데이터에서 광고성 게시글, 사행성 게시글이 상당한 비율을 차지하며, 광고성 글에 대한 필터링 방안의 연구가 병행되어야 제안한 모델의 효율성을 높일 수 있을 것이다. 리트윗, 공유 등으로 중복되는 게시글의 경우 현재 데이터베이스에 저장하는 과정에서 필터링하고 있으나, 리트윗 횟수, 공유 횟수를 검색 과정에서 계산하여 가중치를 줄 수 있을 것으로 보인다. 마지막으로 이미지 데이터 역시 수집하여 각 그림의 메인 색상을 추출해 관련 색상으로 항목을 추가할 수 있다.

ACKNOWLEDGMENT

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 해외 ICT 전문인력활용촉진사업의 연구결과로 수행되었음" (IITP-2016-R0134-15-1030) 지원에 의해 연구되었음.

참고문헌

- [1] Raut, Vijay B., and D. D. Londhe. "Survey on Opinion Mining and Summarization of User Reviews on Web." IJCSIT) International Journal of Computer Science and Information Technologies 5.2 (2014): 1026-1030.
- [2] Nakov, Preslav, et al. "Semeval-2013 task 2: Sentiment analysis in twitter." (2013).
- [3] Hochman, Nadav, and Lev Manovich. "Zooming into an Instagram City: Reading the local through social media." First Monday 18.7 (2013).
- [4] 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 제 26 회 한글 및 한국어 정보처리 학술대회 논문집, 2014.
- [5] Raut, Vijay B., and D. D. Londhe. "Survey on Opinion Mining and Summarization of User Reviews on Web." IJCSIT) International Journal of Computer Science and Information Technologies 5.2 (2014): 1026-1030.
- [6] Mikalai Tsytarau, Themis Palpanas "Survey on mining

subjective data on the web", Data Mining Knowledge Discovery, Springer 2012, pp.478-514.

[7] Chien-Liang Liu, Wen-Hoar Hsiao, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou "Movie Rating and Review Summarization in Mobile Environment", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May 2012, pp.397-406.