LDA 모델을 이용한 트위터 데이터 분석 시스템

이일섭, 장정현, 류관희
충북대학교 소프트웨어학과
e-mail:lis123kr@cbnu.ac.kr

Twitter Data Analysis System using LDA model

Il Seob Lee, Jeong Hyeon Jang, Kwan-Hee Yoo
Dept of Computer Science, Chungbuk National University

요 약

현재 많은 사용자들이 모바일 기기를 통해 소셜 네트워크 서비스(이하 SNS)를 이용하고 있으며, SNS를 통해 수많은 데이터가 생성되고 있다. SNS상의 정보는 다양하고 신속하게 다루어지기 때문에 시대의 주요 사건을 잘 표현한다. 본 논문은 2015년 1월부터 2017년 8월까지의 약 191만개의 트위터 데이터를 수집한 후, LDA 모델링을 통해 주요 키워드를 추출하고 시대별 주요 토픽과 단어를 파악할 수 있는 시스템을 제안한다.

1. 서 론

최근 트위터, 페이스북, 인스타그램과 같은 소셜 네트워크 서비스를 통해 자신의 관심사, 경험을 공유하며 다양한 사람과 소통하려는 사람이 증가하고 있다. 또한, 시대의 주요 사건에 따른 이용자들의 다양한 의견들이 게시된다. 특히, 트위터는 트윗이라는 짧은 길이의 텍스트를 통해 다양한 주제의 게시글이 실시간으로 생성된다. 이러한 트위터 데이터를 이용하여 의미 있는 정보를 찾아내려는 연구가 증가하고 있다[1][2].

본 논문에서는 2015년 1월부터 2017년 8월까지의 약 191만개의 트위터 데이터를 이용해 SNS상에서 이슈가 되는 토픽을 추출하기 위해 LDA 토픽 모델을 이용한다. 이 결과를 통해 시대별 주요 토픽을 확인할 수 있고, SNS상에서 많이 언급되고 있는 주요 키워드를 확인할 수 있는 시스템을 제안한다.

2. 관련 연구

LDA(Latent Dirichlet Allocation) 모델[1]은 주어진 문서에 대하여 각 문서에 어떤 주제들이 존재하는지에 대한 확률 모형으로, 미리 알고 있는 주제별 단어 수 분포를 바탕으로, 주어진 문서에서 발견된 단어수 분포를 분석함으로서 해당 문서가 어떤 주제들을 함께 다루고 있을지를 예측한다. LDA 토픽 모델을 이용해 트위터의 주요 키워드를 추출해 최신 트렌드를 확인하거나 추출된 토픽 단어 그룹의 카테고리를 진단하는 연구가 진행되었다. 또한, 감성분석과 같은 다양한 텍스트 마이닝 알고리즘을 이용하여 SNS상의 다양한 분석이 진행되고 있다[1-4].

3. 트위터 데이터 분석 시스템

트위터 데이터를 분석하기에 앞서 데이터를 수집하고 전처리하는 과정이 필요하다. [그림 1]은 전체 시스템의 구성도를 나타낸 모습이다. 먼저 정규화(normalization)를 통해 글의 각종 초성 및 오타를 수정한다. 정규화 된 글은 토큰화(tokenization)를 통해 품사를 판별하고 어근화(stemming)를 통해 기본형으로 만든 후, 어구 추출(phrase extraction)을 통해 데이터 분석에 필요한 데이터를 보완할 수 있으며, 불필요한 데이터는 제거한다. 이와 같은 전처리 과정은 python의 twkorean 모듈[5]을 이용하였다.

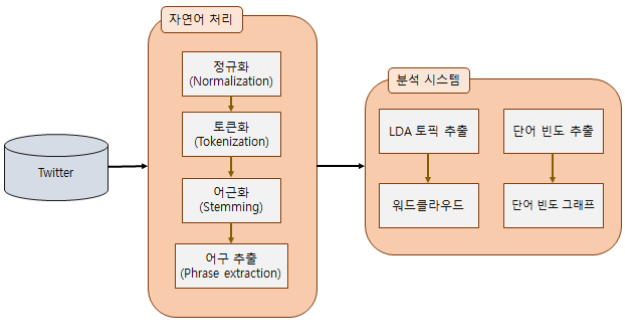

그림 1. 시스템 구성도

전처리된 데이터는 LDA 토픽 모델을 통해 토픽을 추출과 단어빈도를 계산한다. LDA 토픽 모델은 사전에 토픽 그룹 수를 정의해야 한다. 이를 위해 토픽 그룹 수를 조정해가며 실험한 결과 14개의 토픽 그룹 수가 가장 단어들이 유의미한 관계를 갖는 결과를 얻었다. [표 1]은 2017년 7월 약 6만개의 트윗에서 추출된 토픽 그룹 결과이다. 여러 토픽 그룹을 묶어 카테고리를 파악할 수 있다. 그룹 1, 9를 묶어 '정치'라는 카테고리로, 그룹 2, 6, 11을 묶어 '연애'라는 카테고리로 진단할 수 있다.

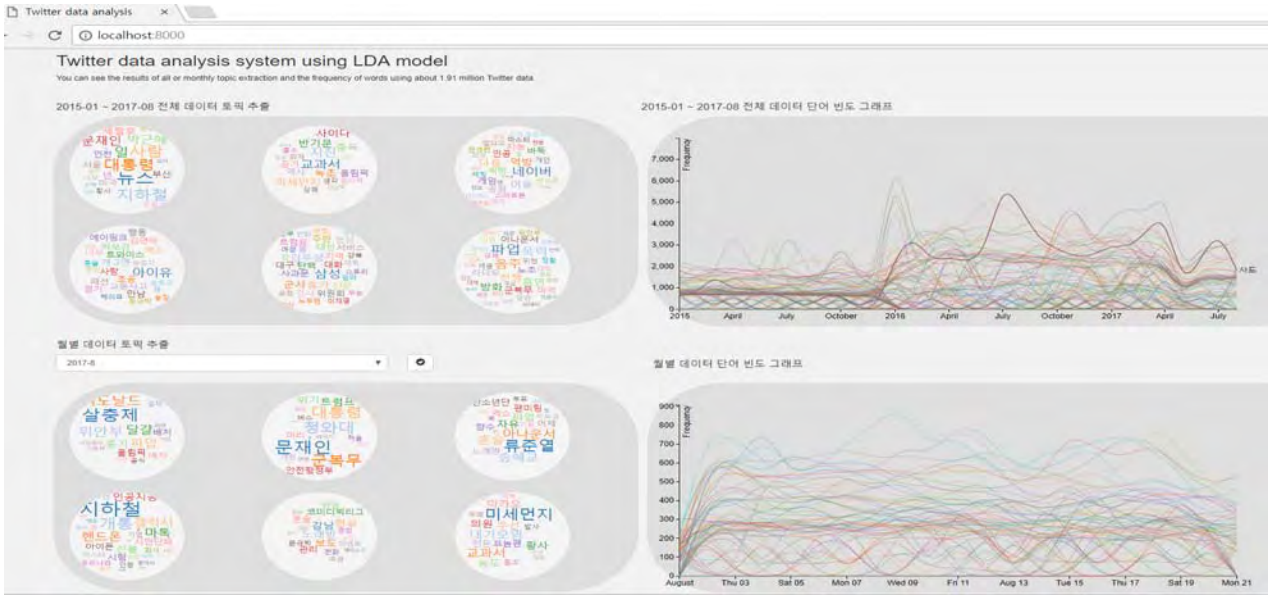

그림 2. 트위터 분석 시스템

표 1. 2017년 8월의 추출된 토픽 그룹

그룹 번호	단어 그룹
1	문재인, 대통령, 청와대, 트럼프, 취임
2	코미디빅리그, 이벤트, 방송, 에피소드
3	프놈펜, 마카오, 향수, 소말리아
4	미세먼지, 대기오염, 황사, 농도
5	핸드폰, 선불, 개통, 지하철, 갤럭시
6	류준열, 송혜교, 아이유, 박보검, 팬미팅
7	올림픽, 간통죄, 강남역, 살충제, 달걀
8	맥도날드, 햄거버, 이벤트, 환타, 병
9	교과서, 정부, 협약, 녀석, 체결
10	스타크래프트, 마스터, 캠페인, 프로토스
11	팬카페, 입국, 청소년, 실검, 사전예약
12	글로벌, 마마, 편집, 컨셉, 노래방
13	외국인, 부산여행, 사랑, 인천
14	제명, 직업, 소액, 제작, 사이버, 결제

본 시스템은 전체 분석과 월별 분석으로 구분된다. 전체 분석은 모든 데이터를 이용한 토픽 추출과 단어 빈도 변화를 보이며, 월별 분석은 사용자가 선택한 연도, 월의 결과를 보여준다. LDA 토픽 모델을 통해 추출한 토픽 그룹은 워드클라우드 시각화 방법을 통해 보인다. 그리고 시간에 따라 빈도가 높은 50개의 단어를 추출해 시간에 따른 단어 빈도 그래프를 보인다. [그림 2]는 제안하는 시스템의 구현 결과이다. 시스템을 통해 전체분석 결과에서는 '정치', '사회', '연예', 'IT/기술' 카테고리로 그룹 된 단어들의 결과를 파악할 수 있으며, 월별분석에서는 2017년 8월에 이슈 된 '사회' 카테고리의 '살충제', '달걀', '맥도날드' 등의 단어가 더 비중 있게 다루어진 모습을 확인할 수 있다.

4. 결 론

빠르고 간편하게 글을 게시할 수 있는 SNS의 특성상 시대의 주요 이슈에 대한 글이 많이 게시된다. 본 논문에서는 약 191만개의 트위터 데이터를 수집하여 SNS상의 주요 이슈가 되는 토픽을 추출하고 단어의 빈도를 확인할 수 있는 시스템을 제안하고 그 결과를 확인하였다.

LDA 모델은 일반 문서와 달리 짧은 글이 많은 SNS의 경우 모델의 성능이 저하될 수 있다. 이러한 문제를 해결하기 위한 LDA 모델의 개선과 함께 한글에 대한 자연어 처리를 개선하야 할 향후 과제가 있다[6].

Acknowledgement

본 논문은 교육부가 지원하고 충북대학교가 수행하는 지역선도대학육성사업의 지원을 받아서 수행되었습니다.

참고문헌

[1] 정병문, 김태환, 이진, 김정선 "LDA 모델을 이용한 트위터 토픽 추출 및 토픽 카테고리 판단", 한국정보과학회 2015년 동계학술발표회 논문집, pp. 787-788, 2015.

[2] 류우종, 하종우, Md. Hijbul Alam, 이상근, "토픽 모델링 기법을 이용한 트위터 트렌드 추출", 한국정보과학회 2013년 추계학술발표회 논문집, pp. 191-193, 2013.

[3] 최홍구, 황인준, "트위터 문서 분석을 통한 감정 기반의 음악 추천 시스템", 정보과학회논문지, Vol. 18, No. 11, pp. 762-767, 2012.

[4] 이루다, 김진만, 임좌상, "LDA를 이용한 트위터의 토픽 분석", 한국통신학회 동계종합학술발표회 논문집, pp. 1010-1011, 2016.

[5] python twitter-korean 0.1.0.dev522, https://pypi.python.org/pypi/twitter-korean/ 0.1.0.dev522.

[6] Wayne Xin ZhaoJing JiangJianshu WengJing HeEe-Peng LimHongfei YanXiaoming Li, "Comparing Twitter and Traditional Media Using Topic Models", ECIR 2011, pp.338-349, 2011.