텍스트 마이닝 사례 소개 및 활용 방안 제안

Introduction of Text Mining Case and Proposal of Application Plan

저자
(Authors)

박지원, 김응모
Park Ji-won, KimUng-mo

출처
(Source)

Proceedings of KIIT Summer Conference , 2017.12, 329-330 (2 pages)

발행처
(Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/Article/NODE07274894

APA Style

박지원, 김응모 (2017). 텍스트 마이닝 사례 소개 및 활용 방안 제안. Proceedings of KIIT Summer Conference, 329-330.

텍스트 마이닝 사례 소개 및 활용 방안 제안

박지원(*), 김응모(**)
(*) 성균관대학교 소프트웨어대학, jihyews123@hanmail.net
(**) 성균관대학교 소프트웨어대학, umkim@skku.edu

Introduction of Text Mining Case and Proposal of Application Plan

Park Ji-won(*), Kim Ung-mo(**)
(*) Sungkyunkwan University, Department of Software
(**) Sungkyunkwan University, Department of Software

요 약

본 논문에서는 여러 빅 데이터 분석 기술 중 하나인 텍스트 마이닝의 절차를 간략히 소개한 후 이를 활용하여 소셜 빅 데이터를 분석한 사례를 연구하였다. 또한, 이를 분석하여 텍스트 마이닝 기술의 활용 방안을 제안한다.

1. 서론

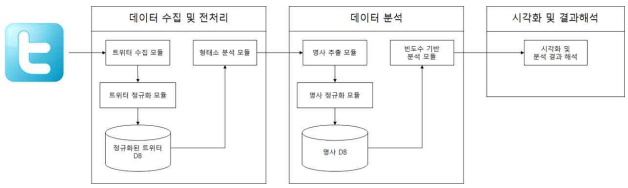
소셜 빅 데이터에는 사용자들의 솔직하고 다양한 의견이 반영되어 있어 진실성과 진정성이 확보된 데이터로서의 높은 가치를 지니고 있다. 이렇게 생성된 '비정형 텍스트'를 가지고 의미 있는 정보를 추출하기 위해서는 텍스트 마이닝을 통한 분석이 필요하다. 텍스트 마이닝은 구조화되지 않은 대규모의 데이터를 수집하고 분석하여 숨겨져 있는 유용한 정보를 찾아내는 과정을 의미한다.

본 논문에서는 텍스트 마이닝을 적용한 사례를 탐색하고 이를 분석하여 보완점과 활용 방안을 제시한다.

2. 텍스트 마이닝 기법 절차

트위터를 통한 텍스트 마이닝 과정은 다음과 같다.

(그림 1) 텍스트 마이닝 과정



(Figure 1) Process of Text Mining

2.1 데이터 수집 및 전처리

트위터 API를 이용해 데이터를 수집하고 전처리 과정을 통해 컴퓨터가 처리하기 쉽도록 정형화한다. 사전에 정의되지 않은 텍스트 또는 기호들을 필터링한 후 R의 tm, KoNLP 패키지를 활용하여 자연어 처리 과정을 거친다.

2.2 데이터 분석

전처리 과정을 거친 용어들은 데이터베이스로 구축된다. 수집된 데이터를 가지고 핵심 키워드와 연관성을 분석한다.

2.3 시각화 및 결과 해석

분석한 결과를 워드 클라우드, 단어 기반 계층적 클러스터링, k-means 클러스터링, k-medoid 클러스터링 등의 방법으로 시각화한다.

3. 텍스트 마이닝 적용 사례 및 활용 방안

3.1 영화 <변호인>의 흥행 요인 분석

(1) 개요 : 영화가 개봉되기 전 첫 주의 소셜 빅 데이터는 영화의 흥행 수익과 강하게 연관되어 있다고 보고 있다. 앞으로 빅 데이터 분석을 통한 영화 흥행 예측이나 요인 분석 결과를 지표로 활용하여 마케팅 전략을 수립하려는 사례가 늘어날 것으로 전망된다.

(2) 실험 과정 : 영화 <변호인>이 개봉하기 전인 2013년 11

월 1일부터 2014년 3월 1일까지의 트윗을 수집하여 영화의 흥행 요인을 분석하였다. 크롤러를 이용하여 선정된 키워드를 포함한 트윗을 수집하였다. 전처리 과정과 형태소 분석을 통해 명사를 추출하여 단어들의 빈도수를 계산하였다.

(3) 분석 결과 : 영화 <변호인>의 흥행에 긍정적인 영향을 미친 속성으로 '스토리', '배우', '감독'을 뽑을 수 있었다.

(4) 향후 활용 방안 제안 : 다양한 장르의 영화에 텍스트 마이닝을 적용하여 장르별 흥행 요인을 파악하는 것은 영화 산업에서 중요한 정보원으로 활용될 것이다. 또한 관객들이 선호하는 요인을 미리 예측함으로써 제작 과정에서도 영화 흥행의 가능성을 높일 수 있을 것으로 기대된다.

3.2 2012년 한국대선 관련 트윗 분석

(1) 개요 : 소셜 미디어를 분석하여 선거 결과를 예측하고자 하는 연구는 전 세계적으로 활발하게 진행되고 있다. Williams와 Gulati의 연구 결과에 의하면 소셜 미디어에서 나타나는 지지율은 유력한 후보자를 예측할 수 있는 지표로 사용될 수 있다.

(2) 실험 과정 : 2012년 10월 1일부터 10월 31일까지 '박근혜', '문재인', '안철수', '대선'이라는 단어가 포함된 트윗을 수집하였다. 수집한 데이터를 가지고 세 후보의 이름을 질의어로 사용하여 동시 출현 단어를 도출하였다. Mimno와 McCallum가 제안한 다항 토픽 모델링 기법을 이용해 결과를 분석하였다.

(3) 분석 결과 : 논쟁이 되는 이슈가 신문기사보다 소셜 미디어에서 더욱 빠르게 전파되고 있음을 확인할 수 있었다. 또 신문기사는 세부 사건 변화 관찰에, 트위터는 그에 대한 영향력 변화 관찰에 적합함을 확인할 수 있었다.

(4) 향후 개선 방안 제안 : 연구를 위해 수집한 데이터는 '대선 후보의 이름'이라는 질의어만 가지고 추출한 것이기 때문에 배제된 데이터가 존재할 것이다. 따라서 이후 연구에서는 해당 이슈에 대한 흐름을 더 자세히 파악하여 놓친 정보를 포함시킬 수 있는 방법이 논의되어야 한다. 또한, 감정 분석 등을 추가적으로 수행한다면 여론의 변화 과정을 더 심도 있게 관찰할 수 있을 것으로 기대된다.

3.3 사용자의 감정에 적합한 음악 추천

(1) 실험 과정 : Thayer가 제안한 11개의 무드 중 angry, sad, happy, peacful 4개의 무드를 기준으로 실험을 진행한다. 트위터에서 사용자의 감정을 추출하기 위해 ANEW와 WordNet을 이용하여 단어뭉치를 작성한다.

멀티 모달 무드 분류기를 사용하여 주어진 음악 데이터베이스를 분석하고, 5명의 피실험자를 대상으로 트위터에서 얻은 감정과 일치하는 무드의 음악을 무작위로 추천한다.

(2) 분석 결과 : 만족도 조사를 수행한 결과 음악을 추천했을 때는 평균 6.8의 만족도를, 음악 추천이 없을 때는 평균 5.2의 만족도를 얻었다. 즉, 음악 추천이 사용자에게 감정적인 만족을 주는 것을 확인할 수 있었다.

(3) 향후 개선 방안 제안 : 만족도 조사 과정에서 피실험자의 표본 수는 5로 너무 작았다. 결과의 정확도를 높이기 위해선 실험에 참여하는 표본의 수가 커져야 한다. 또 사용자의 감정을 추출할 때, 이모티콘 등 텍스트 외의 데이터에 대한 추가 분석을 통해 정확도를 높일 수 있다.

4. 결론

소셜 미디어의 풀이 커짐에 따라 이를 분석하여 유의미한 결과를 도출해내는 것이 중요해졌다. 이미 많은 기업들은 SNS의 빅 데이터를 수집하고 분석하여 고객 지향적인 아이디어를 내고 있다. SNS상에 존재하는 정보의 흐름을 분석하여 미래 시장을 예측하거나 고객의 니즈를 파악하는 등 빅 데이터로 인해 시장이 받는 영향력은 앞으로도 더욱 커질 것이다. 따라서 향후 텍스트 마이닝의 활용에 대한 전망이 기대된다.

참고 문헌

[1] McKinsey Global Institute, "Big Data: The Next Frontier for Innovation, Competition, and Productivity", www.mckinsey.com/mgi, 2011.05.

[2] http://www.rdatamining.com/examples/text-mining

[3] 이오준, 박승보, 정다울, 유은순, "소셜 빅데이터를 이용한 영화 흥행 요인 분석", 한국콘텐츠학회논문지, 2014.10.

[4] 배정환, 손지은, 송민, "텍스트 마이닝을 이용한 2012년 한국대선 관련 트위터 분석", 지능정보연구, p.141-156, 2013.09.

[5] 최홍구, 탁윤식, 황인준, "트위터 분석을 통한 음악 추천 시스템", 한국정보과학회 학술발표논문집, 2011.11.