Movie Box-office Analysis using Social Big Data

Article · October 2014
DOI: 10.5392/JKCA.2014.14.10.527

4 authors, including:

O-Joun Lee
Chung-Ang University
44 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Eunsoon You
Dankook University
26 PUBLICATIONS **49** CITATIONS
SEE PROFILE

http://dx.doi.org/10.5392/JKCA.2014.14.10.

소셜 빅데이터를 이용한 영화 흥행 요인 분석

Movie Box-office Analysis using Social Big Data

이오준*, 박승보**, 정다울*, 유은순**
단국대학교 소프트웨어학과*, 단국대학교 미디어콘텐츠연구원**

O-Joun Lee(concerto_grs@naver.com)*, Seung-Bo Park(molaal@naver.com)**,
Daul Chung(ekdnfsla@gmail.com)*, Eun-Soon You(tesniere@naver.com)**

요약

수요 예측은 영화 산업에서 매우 중요한 문제이다. 최근 들어 트위터(Twitter), 페이스북(Facebook)과 같은 소셜미디어의 비정형 텍스트 데이터를 이용하여 영화 흥행을 예측하고 분석하는 시도들이 활발하게 이루어지고 있다. 기존에는 주로 데이터의 주기별 변화량을 측정하여 데이터 양과 영화 흥행간의 상관성을 분석하거나 데이터에 대해 감성의 극성 값을 부여하는 오피니언 마이닝을 통해 영화의 흥행 추이를 예측하였다. 하지만 이러한 정량적 접근만으로는 관객들이 영화를 선택하게 된 근거나 영화의 어떤 속성을 선호하는지를 알 수 없기 때문에 영화의 흥행 요인을 밝히는데 한계가 있었다. 따라서 본 연구는 트위터 데이터를 수집한 후 빈도수 측정을 통해 트윗의 내용을 대표하는 토픽(topic) 키워드를 추출하여 관객들의 관심을 반영하는 영화적 속성들이 무엇인지를 밝히고, 그 속성들에 대한 관객들의 반응을 분석함으로써 영화의 흥행에 영향을 미친 요인들을 제시한다.

■ 중심어 :□빅데이터□소셜미디어□영화 흥행 예측□트위터□토픽□

Abstract

The demand prediction is a critical issue for the film industry. As the social media, such as Twitter and Facebook, gains momentum of late, considerable efforts are being dedicated to prediction and analysis of hit movies based on unstructured text data. For prediction of trends found in commercially successful films, the correlations between the amount of data and hit movies may be analyzed by estimating the data variation by period while opinion mining that assigns sentiment polarity score to data may be employed. However, it is not possible to understand why the audience chooses a certain movie or which attribute of a movie is preferred by using such a quantitative approach. This has limited the efforts to identify factors driving a movie's commercial success. In this regard, this study aims to investigate a movie's attributes that reflect the interests of the audience. This would be done by extracting topic keywords that represent the contents of Twits through frequency measurement based on the collected Twitter data while analyzing responses displayed by the audience. The objective is to propose factors driving a movie's commercial success.

■ keyword :□Big Data□Social Media□Movie Box Office□Twitter□Topic□

* 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2013R1A1A2057943)
접수일자 : 2014년 08월 19일
수정일자 : 2014년 09월 11일
심사완료일 : 2014년 09월 11일
교신저자 : 유은순, e-mail : tesniere@naver.com

I. 서 론

모바일 기기의 확산과 소셜미디어의 활성화로 데이터가 폭발적으로 증가하고 있다. 데이터의 양이 현재 테라바이트에서 향후에는 페타바이트, 엑사바이트에 달할 것으로 예상되는 가운데, 기존의 일반적인 기술로 수집, 저장, 관리, 분석하기 어려운 대규모의 데이터를 빅데이터[1]로 정의하고 있다. 현재 폭증하고 있는 데이터의 대부분은 트위터, 페이스북 등과 같은 소셜미디어를 통해 사용자들이 생성하고 있는 비정형 텍스트이다. 이 텍스트 속에는 개인의 일상부터 사회, 정치, 경제, 문화에 대한 사용자들의 솔직하고 다양한 의견과 감성 등이 반영되어 있어서 "진실성과 진정성이 확보된 데이터로 높은 가치[2]"를 지니고 있다.

소셜 빅데이터 분석은 다양한 형태의 소셜미디어에서 실시간으로 생성되는 대규모의 데이터를 수집하고 분석하여 통찰을 이끌어내는 것[3]을 의미한다.

소셜 빅데이터는 시간에 따른 사용자들의 의견과 감정의 변화를 추적하거나, 사건과 이슈를 감지하는데 유용하다. 특히 높은 근접성과 간결성을 특징으로 하는 트위터는 140글자 메시지인 트윗(tweet)을 전달하는 리트윗(retweet) 기능과 단방향 관계를 통해 정보를 확산시킬 수 있기 때문에 실제 세계에서 일어나고 있는 사건을 실시간으로 인식할 수 있는 구조적 장점을 갖고 있다.

최근 들어 많은 분야에서 소셜미디어에 올라오는 의견을 분석하고 의미를 추출하여 이를 활용하려는 시도들이 활발하게 이루어지고 있다. 예를 들어 기업들은 분석을 통해 자사 이미지와 제품에 대한 여론을 분석하고 소비자의 니즈를 발견하여 신속한 대응을 하는 등 기업의 리스크 관리와 경쟁력 확보를 위한 중요한 도구로 삼고 있다. 또한 선거 기간 동안 트위터의 데이터를 분석하여 선거의 흐름과 판도를 예측하기도 한다. 영화 산업도 예외가 아니다. 영화는 소셜미디어에서 광범위하게 언급되는 문화콘텐츠로써 소셜 분석의 좋은 주제가 될 수 있다. 영화 산업에서의 소셜미디어 활용 연구는 SNS에 올라오는 영화 관련 데이터를 이용한 영화 흥행 예측[4-12]과 영화 마케팅에서의 소셜미디어 활용[13-16], 그리고 SNS 데이터에 내재된 개인 성향 분석을 통한 맞춤형 영화 추천 시스템[17-19]으로 정리될 수 있다. 이 중에서 영화 흥행 예측은 가장 많이 연구되고 있는 분야이다.

그동안 영화의 수요 예측을 위해 영화 관련 데이터의 양적 변화를 분석하거나 데이터에 대한 감성의 극성을 결정하는 오피니언 마이닝 연구가 주를 이루었다. 하지만 소셜미디어를 통해 관객들이 생산하는 텍스트에는 맞춤법 오류가 많고 신조어와 은어, 이모티콘 등과 같은 인터넷 용어를 많이 사용하기 때문에 감성 극성 결과의 정확도는 높지 않다. 또한 감성 극성 결과만으로는 관객들의 영화 선택 동기나 선호하는 영화의 속성이 무엇인지 알 수 없다. 영화의 속성은 영화를 구성하는 다양한 요소들로써, 스토리, 감독, 배우, 음악, 특수효과, 편집 등 매우 다양하며, 영화의 장르와 내용에 따라 관객들이 선호하는 속성은 다르게 나타난다.

본 연구는 2013년 12월 18일에 개봉했던 영화 <변호인>의 트위터 데이터를 분석하여 관객들이 영화의 어떤 속성을 선호하는지를 살펴봄으로써 영화 흥행의 주요 요인을 밝히는 것을 목적으로 한다. 부림사건과 고 노무현 대통령을 모티프로 한 이 영화는 정치적 소재로 인해 개봉 전부터 사회적 논란과 화제를 불러 일으키며 개봉 33일만에 천만 관객을 돌파하였고 트위터를 통해 활발한 담론이 형성되었다. 따라서 관객들의 영화 선택이 무엇에 근거하여 이루어졌는지를 분석하여 영화 흥행의 요인을 연구하는데 적합하다고 판단하였다.

영화 <변호인>의 흥행 요인을 분석하기 위해 본 연구는 수집된 트윗을 분석하는 시스템을 구현하여 형태소 분석을 통해 고빈도 명사를 추출하였다. 명사만을 고려한 이유는 명사는 텍스트에서 가장 높은 빈도를 보이는 품사로써 중요한 정보를 전달하기 때문이며[20], 문장에서 명사만으로도 내용의 핵심을 파악하는 것이 가능하기 때문이다[21]. 본 연구에서는 트위터에서 빈번하게 출현한 명사를 토픽으로 정의하였다. 트위터 분석에서 토픽은 "화제가 되는 키워드 혹은 화제가 되는 키워드와 관련된 이야깃거리[21]"를 의미한다. 높은 빈도를 보이는 토픽들은 관객들이 영화에 대해 관심을 가지는 대상이 무엇인지를 보여주기 때문에 관객이 영화

를 선택하는 근거로 가정하였다.

본 논문의 구성은 다음과 같다. II장에서는 국내외 영화 흥행 연구를 살펴보고, III장에서는 소셜 빅데이터를 활용한 영화 흥행 연구 현황을 기술한다. IV장에서는 트윗의 분석 시스템을 소개하고 V장에서는 영화 <변호인>의 흥행 요인을 분석한다. 그리고 마지막 VI장에서는 연구의 결과와 향후 연구를 제시한다.

II. 관련 연구

영화 산업이 지속적으로 팽창하면서 영화 흥행 연구에 대한 관심도 증가하고 있다. 영화는 경험재라는 속성 상 수요 예측이 힘든 문화 상품이다. 따라서 위험 요소를 피하고 수익을 증대시켜야 하는 영화 배급사와 투자자들에게 있어 영화 흥행은 매우 중요한 이슈라고 할 수 있다. 그동안 영화 수요 예측을 위해 제작비, 감독과 배우, 마케팅과 구전, 개봉시기와 스크린 수 확보, 장르 등과 같은 영화 메타데이터들을 주로 사용해 왔다. 하지만 최근에는 SNS가 영화 흥행의 중요한 변수로 인식되면서 SNS의 비정형 텍스트 데이터 분석을 통해 영화 흥행의 추이를 예측하고 분석하는 연구들이 활발하게 이루어지고 있다.

1. 국외 연구

Asur & Huberman(2010)은 2009년도 11월부터 2010년 2월까지 3개월 동안 미국에서 개봉된 영화 24편에 대한 약 290만 건의 트위터 메시지를 분석하여 긍정과 부정, 중립으로 감성을 분류하였다. 그리고 영화 개봉 전 일정 기간에 트위터에서 언급된 영화 횟수와 개봉 첫 주의 영화 수입 간의 유의미한 상관성을 밝혀냈다. 영화 24편에 대한 개봉 첫 주 수입 예측 정확도는 97.3%로 Hollywood Stock Exchange(HSX)의 정확도 96.5% 보다 높은 것으로 나타났다[11]. Lica & Tuta(2011) 역시 트윗의 감정 분석을 통해 영화 30편에 대한 흥행 성적을 예측하고 실제 인터넷 무비 데이터베이스(IMDB)의 박스 오피스 결과와 비교하였다[9]. Mishne & Glance(2006)는 영화 개봉 전 웹 블로그 상에서 유저들이 포스팅한 내용에 대한 감정 분석을 통해 긍정적 감정이 영화 흥행을 예측하는 중요한 변수임을 증명하였다[8].

SNS의 데이터 뿐만 아니라 영화 리뷰를 이용하여 영화 흥행을 예측하는 연구 사례들도 있다. Doshi et al.(2010)은 소셜 네트워크 분석과 감정 분석을 결합한 웹 마이닝 기술을 통해 인터넷 영화 데이터베이스(IMDB)의 영화 리뷰를 분석하고 영화의 흥행 여부를 예측하였다[22]. Joshi et al.(2010)은 영화 리뷰에 대한 감정 분석을 통해 영화의 오프닝 주말 수입을 예측하였다[10].

텍스트의 감성을 분석하는 오피니언 마이닝 이외에도 방대한 양의 트윗에서 의미 있는 정보를 추출하기 위해 토픽을 이용하는 연구들이 진행되고 있다. Kim et al.은 코아 토픽 기반의 클러스터링(Core-Topic-based Clustering)을 이용하여 특정 인기 TV 드라마의 트윗을 분석하고 의미 있는 토픽들을 추출하였다[23]. Ni et al.은 트위터에서 핵심이 되는 코아 단어(core word)를 추출하여 사용자들에 의해 많이 언급된 영화 내용이 무엇인지를 제시하고 토픽에 따라 유사한 트윗을 클러스터 하였다[24].

2. 국내 연구 사례

강지훈, 박찬희, 도형록, 김성범은(2014) 영화의 메타데이터뿐만 아니라 평점, 댓글 수와 같은 관객 반응을 고려한 변수들을 함께 이용하여 영화의 수요를 예측하고 실제 영화 흥행 성적과 비교하였다[4]. 허민회, 강필성, 조성준(2013) 역시 영화 관련 정보와 네이버의 영화 평점에 대한 의미 분석을 통해 영화에 대한 관객의 긍정과 부정 표현의 발현 정도가 영화 흥행과 어떠한 상관성이 있는지 제시하였다[25]. SNS 데이터를 이용한 영화 흥행 예측 연구도 활발히 진행되고 있는데, 김진욱(2014)은 영화 홍보용 SNS에 대한 대량의 검색 조회수를 이용하여 영화 <설국열차>와 <더 테러 라이브>의 흥행 예측 결과를 비교하였다[7].

박선영(2012)은 영화 개봉 전과 개봉 초기, 성숙기로 기간을 나누고 영화<써니>의 기간별 SNS 메시지 변화량을 분석하여 SNS를 통한 구전 효과와 영화 흥행간

의 상관성을 증명하였다[26]. Kim et al.(2013)은 회귀분석과 확산모델(Bass diffusion model)을 이용하여 페이스북과 트위터의 메시지를 분석하여 영화의 메타데이터 보다 SNS의 긍정 및 부정적 언급이 영화 흥행에 더 중요한 영향을 미치는 변수임을 제시하였다[12].

지금까지 진행된 국내외 선행연구들을 살펴보면 영화에 대한 관객의 선호도와 흥행의 추이를 예측하기 위해 SNS의 비정형 데이터를 분석하여 감성의 극성을 분류하거나 데이터의 시계열적 변화를 추적하는 정량적 관점이 주를 이루고 있다. 하지만 언어 표현의 풍부함과 복잡한 내재적 속성 때문에 텍스트가 함축하고 있는 감성을 정확하게 식별하고 극성 값을 부여해야 하는 감성 분석은 매우 어려운 작업이다. 특히 문맥을 고려해야 하는 다의적(多義的) 표현이나 유머, 풍자, 아이러니와 같은 복잡한 표현을 분석할 때 오류가 발생하기 쉽다. 텍스트에 대한감성의 극성 분류는 중요하고 유용한 문제이지만 감성 극성의 결과만으로는 사용자들이 무엇을 좋아하고 싫어하는지는 발견할 수 없다. 따라서 본 논문은 정량적 분석 결과에 근거하여 트위터의 내용을 분석하는 정성적 분석을 통해 관객들이 영화의 어떤 속성을 선호하는지를 살펴봄으로써 영화 흥행의 주요 요인을 밝혀보고자 한다.

III. 소셜 분석 솔루션을 이용한 영화 흥행 연구 현황

영화는 소셜미디어 사용자들 사이에서 가장 많이 이야기되는 대중문화 중의 하나로 소셜미디어에는 영화와 관련된 방대한 양의 데이터가 존재한다. 특히 개봉을 앞두거나 개봉 직후의 대형 블록버스터 영화들은 소셜미디어에서 관객들의 폭발적인 관심과 반응을 이끌어내며 인기 토픽으로 등장한다. 그리고 영화 시사회나 개봉 후에 소셜미디어를 통해 생산되고 공유되는 정보들은 관객들의 자발적인 구전을 이끌어 냄으로써 영화 흥행에 긍정적인 영향을 미친다. 특히 영화가 개봉되기 전 첫 주의 소셜 빅데이터가 영화의 흥행 수익과 가장 강하게 연관되어 있다고 보고 있다. 이처럼 소셜미디어와 영화 흥행 수익 간에 강한 상관성이 존재함에 따라 영화 산업에서 소셜 빅데이터는 영화에 대한 관객의 반응과 변화를 관찰하고 영화의 흥행 추이를 예측하는데 중요한 정보원으로 활용되고 있다. 실제로 할리우드 영화사들은 영화가 개봉되기 전과 후에 관객의 반응을 추적하고 흥행 실적을 예측하는 서비스를 영화 마케팅에 적극적으로 이용하고 있다.

소셜미디어 여론 분석 업체 피지올로지(Fizziology)는 2009년 서비스 개시 이후 500개 이상 영화의 박스오피스 실적을 추적해왔다. 영화가 개봉되기 4주 전부터 개봉 후 3주 기간에 걸쳐 트위터, 페이스 북 등의 API를 이용하여 데이터를 수집하고 분석하여 영화 흥행 예측 서비스를 제공하여 영화사가 효과적인 마케팅 전략을 전개할 수 있도록 지원하고 있다. [그림 1]은 피지올로지가 트위터와 페이스북에서 언급된 영화 <그래비티>의 데이터를 분석한 예시 화면이다[27].

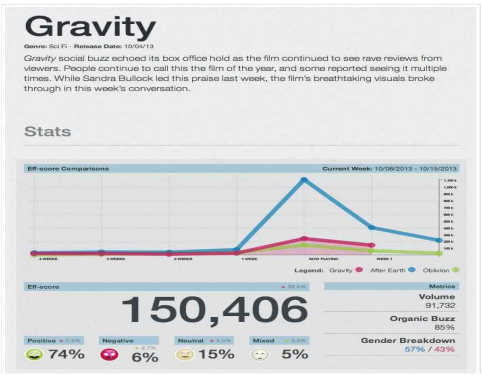

그림 1. 영화 <그래비티>에 대한 소셜빅데이터 분석 화면

국내에서도 다양한 소셜 분석 솔루션들이 활발하게 개발되고 있는 가운데, 메조미디어에서 개발한 티버즈(TIBUZZ, Talking Index Buzz)는 흥행에 성공한 영화 44편을 선별하고 소셜미디어에서 해당 영화에 대해 언급한 데이터 70만건 이상을 분석한 결과, 소셜미디어에 올라온 버즈(buzz)량과 영화 관객수가 비슷한 흐름을 보이고 있음을 밝혀내었다[28][29]. [그림 2]는 영화 44편에 대한 개봉 전후 일주일간의 총버즈량과 총관객수의 증감을 비교한 것이다. 특히 개봉직전 보다 개봉 직후의 버즈량과 영화 흥행 사이에 더 밀접한 상관관계를

보이는 것으로 나타났다.

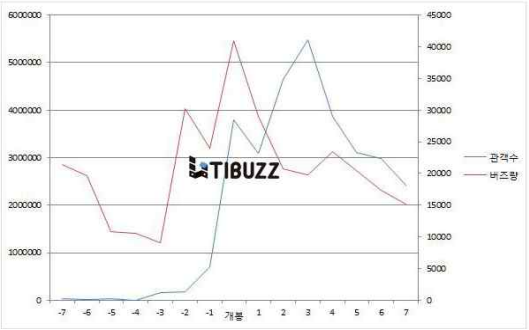

그림 2. 버즈량과 관객수간의 비율을 보여주는 서비스 화면

코난 테크놀로지의 소셜 분석 서비스인 펄스 K 역시 트위터, 페이스북, 미투데이 등에서 생산되는 수백 만 건의 데이터를 분석하여 영화의 인지도와 호감도를 점수로 계량화함으로써 영화 흥행 예측에 활용하고 있다. 예를 들어 2013년도 1월에 개봉한 한국 영화 <박수건달>, <7번방의 선물>, <베를린>에 대한 트위터 버즈량의 변화 양상을 분석한 결과, 천만 관객을 돌파한 영화 <7번 방의 선물>은 [그림 3]에서도 보이는 것처럼 개봉 전에는 트위터 언급 점유율이 두 영화에 비해 제일 낮았다. 하지만 개봉 후에는 두 영화보다 크게 높아지면서 관객 수도 트위터 언급 점유율과 비슷한 흐름을 보였다[28]. 그림은 개봉되기 전 영화의 트위터 언급 점유율을 나타낸 것이다[30].

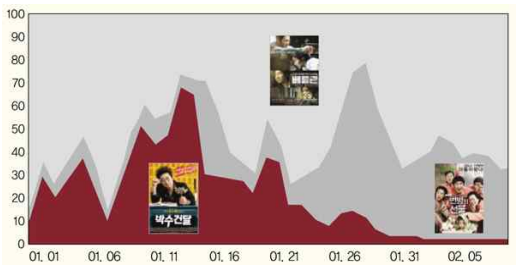

그림 3. 영화 3편의 트위터 언급 점유율

앞으로 빅데이터 분석을 통한 영화 흥행 예측이나 요인 분석 결과를 영화 산업 전반에 활용하려는 사례가 늘어날 것으로 전망된다.

IV. 실험 및 시스템 구현

본 연구는 영화 흥행 요인을 분석하기 위해 영화 <변호인>의 트윗을 이용하였다. [그림 4]는 트윗 수집 및 분석 시스템 구조도로 트윗을 수집하고 전처리하는 파트와 트윗을 분석하는 파트로 구성되었다.

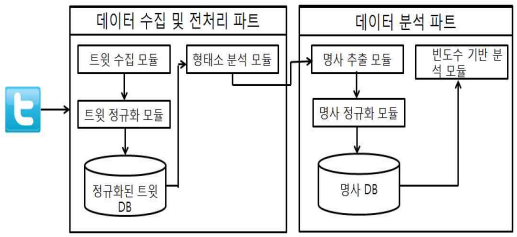

그림 4. 트윗 수집 및 분석 시스템 구조

1. 데이터 수집 및 전처리 파트

관객들은 화제작들이 개봉되기 전부터 SNS를 통해 영화에 대한 뜨거운 관심과 반응을 표출한다. 따라서 본 연구는 영화<변호인>이 개봉되기 전인 2013년 11월 1일부터 2014년 3월 1일까지 트위터에 작성된 영화 관련 메시지를 분석 대상으로 하였다. 트윗 수집을 위해 영화 관련 키워드 11개를 선정하고, 크롤러[31]를 이용하여 키워드를 포함한 91706개의 트윗을 수집하였다. 키워드는 네이버 포털의 '실시간 급상승 검색어'에 근거하여 선정하였다. 네이버의 '실시간 급상승 검색어'는 네이버 검색창으로 입력되는 검색어 중 입력 횟수의 증가 비율이 가장 큰 검색어를 순서대로 보여주는 서비스이다. [표 1]은 선정된 영화 관련 11개의 키워드를 나타낸 것이며, '부산의 학림 사건'과 '부림 사건'처럼 동일한 의미를 가진 키워드들은 하나로 통합하였다.

표 1. 영화 <변호인>과 관련된 11개 키워드

영화 <변호인>과 관련된 11개의 키워드
영화 변호인, 부림 사건(=부산의 학림 사건), 노무현 인권 변호사, 송강호, 송우석 노무현, 양우석, 평점테러, 티켓 테러(=좌석 테러), 최병국 검사

수집된 트윗을 검토한 결과 '부림 사건', '노무현', '평점테러' 등과 같이 정치적 키워드를 포함한 트윗의 대

부분은 영화 자체에 대한 감정과 의견 보다는 정치적 논쟁이나 역사적 사건에 대한 정보를 소개하고 있었기 때문에 본 실험에서 제외했다. 영화 외적인 내용을 언급하는 트윗을 제외하고 영화에 대한 관객의 평가를 포함한 트윗은 총 51253개이며, 이중 내용 중복 제거를 통해 총 30032개의 정제된 트윗을 선별하여 DB로 구축하였다.

선별된 트윗의 전처리 과정으로 링크 URL, 이모티콘, 'ㅠㅠ', 'ㅋㅋ'등과 같은 인터넷 용어를 제거한 후 KAIST CILab에서 개발한 한나눔 한국어 형태소 분석기[32]를 이용하여 형태소 분석을 수행하였다.

2. 데이터 수집 및 데이터 분석 파트

2.1 명사 추출 및 저장

형태소 분석을 통해 명사를 추출 한 후 추가적으로 실제 분석에 필요 없는 영어, 숫자와 같은 노이즈들을 제거하였고, 같은 의미를 지닌 다양한 표현들도 하나로 통합하였다. 예를 들어 '스토리'와 '이야기', '시나리오'는 '스토리'로, '감독'과 '양우석 감독'은 '감독'으로 통합하였다. 전처리 과정을 거친 명사들은 DB로 구축되었다. [그림 5]는 한나눔 형태소 분석기를 이용하여 명사 추출을 실행한 결과를 보여준다. 그림에서 보는 것처럼 형태소 분석을 통해 나온 명사는 '/N'이라는 품사 태깅으로 표시된다.

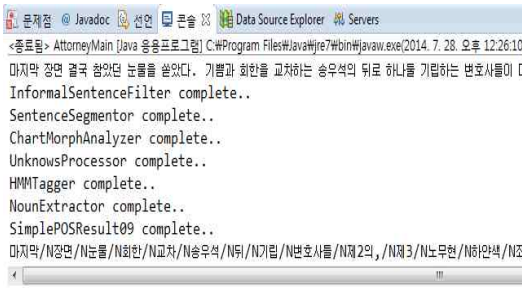

그림 5. 형태소 분석을 통한 명사 추출 결과

'/N' 직전까지의 데이터를 하나의 단어로 간주하고, 자바의 StringTokenizer를 이용하여 문장을 처음부터 읽어나가면서 '/N'을 기준으로 단어들을 토큰화하여 분리하였다. 예를 들어 "오늘 변호인을 보았는데 영화가 끝나고 관객들이 모두 박수를 쳤다"라는 문장의 경우, '변호인', '영화', '관객들', '박수'가 토큰화 되어 추출된다.

2.2. 빈도수 기반의 토픽 추출

토큰화(Tokenization) 과정을 거친 후 단어들의 빈도수를 계산하기 위해 자바의 자료구조 중의 하나인 TreeMap[33]을 사용하였다. 모든 단어를 입력한 후 정렬 메소드로 count 값이 가장 높은 순서대로 출력하였다. 위에서 기술했듯이, 트위터 분석에서 토픽은 사용자들의 관심과 흥미를 반영한 키워드이기 때문에 방대한 양의 트윗의 내용을 직관적으로 이해하는데 유용하다. 따라서 본 연구에서는 트위터에서 빈번하게 출현한 고빈도 명사들은 관객들이 영화에 대해 관심을 갖고 있는 대상이 무엇인지를 보여주고 있기 때문에 토픽 키워드로 간주하였다. 표는 1위부터 20위에 해당하는 토픽들을 나타낸 것이다.

표 2. 빈도수 기반의 토픽 추출

순위	명사	빈도수
1	송강호	17852
2	노무현	11236
3	돌파	4195
4	국민	3815
5	부림사건	1369
6	배우	1104
7	흥행	972
8	눈물	955
9	연기	891
10	대한민국	822
11	이야기	730
12	관람	718
13	감독	737
14	감동	685
15	천만	523
16	장면	474
17	현실	471
18	시대	457
19	임시완	411
20	국가	391

가장 높은 빈도수를 보인 토픽은 영화 제목인 '변호인'과 '영화'이다. 이 두 개의 토픽은 의미 없는 키워드이기 때문에 연구에서 제외되었다. 이 두 개의 키워드를 제외하고 가장 높은 빈도를 보인 토픽은 '송강호'이

다. ‘송강호’이외에 ‘배우’, ‘임시완’, ‘연기’, ’감독‘도 빈번하게 언급됨으로써 관객들은 영화 속 배우와 송우석 감독에 대해 높은 관심을 보이고 있음을 알 수 있다. 영화의 소재가 된 ‘고 노무현 대통령’과 ‘부림사건’, 그리고 영화의 속성을 나타내는 ‘이야기’, ‘장면’도 높은 빈도를 보였다. 또한 ‘눈물’, ‘감동’과 같은 감정 어휘를 비롯하여 ‘흥행’, ‘최고’, ‘감사’, ‘천만 관객’과 같이 영화에 대한 긍정적인 의견을 나타내는 단어들도 포함되어 있었다. ‘대한민국’, ‘현실’, ‘시대’, ‘국가’와 같은 토픽들이 높은 빈도를 보인 이유는 영화 속 배경이 되는 80년대와 2000년대의 한국 현실이 오버랩되면서 관객들로부터 깊은 공감을 이끌어 낸 것으로 보인다.

3. 분석 시스템 구현

DB구축을 위해 Mysql를 사용하였고, [그림 6]에서와 같이 선별된 30032개의 트윗을 입력하는 ‘twitter 테이블’과 트윗의 형태소 분석 결과를 입력하는 ‘morpheme 테이블’을 각각 구성하였다. 30032개의 트윗은 Toad for Mysqul 워크벤치 툴을 사용하여 twitter 테이블에 입력 및 저장되었다. 그리고 각각의 테이블은 number, id, date, 그리고 트윗의 내용을 입력하는 content의 영역으로 구성되었다.

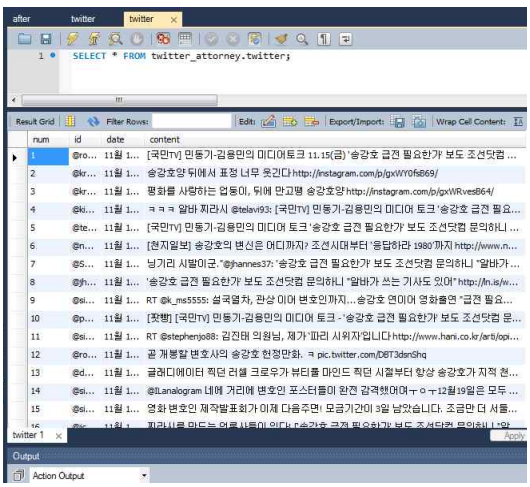

그림 6. 30032개의 트윗 DB

시스템 개발은 자바 이클립스 환경에서 이루어졌는데, 위에서 소개한 트윗 DB와 한나눔 형태소 분석기를 자바 이클립스와 연동하였다. [그림 7]과 [그림 8]은 각각 형태소 분석 기반의 명사 추출 결과와 명사 빈도수 결과를 보여주고 있다.

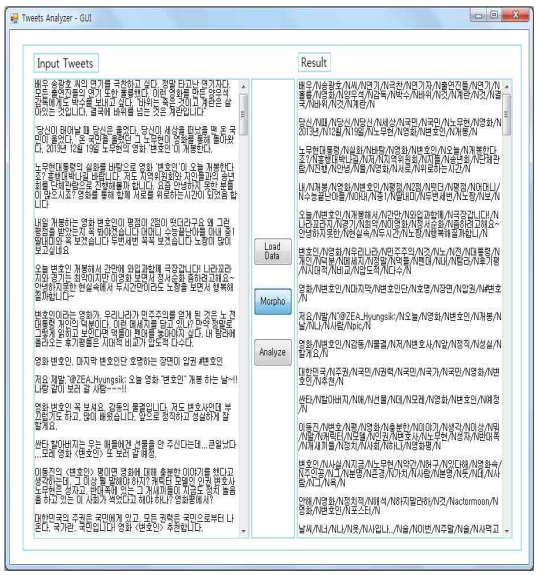

그림 7. 시스템의 형태소 분석 기반 명사 추출 결과

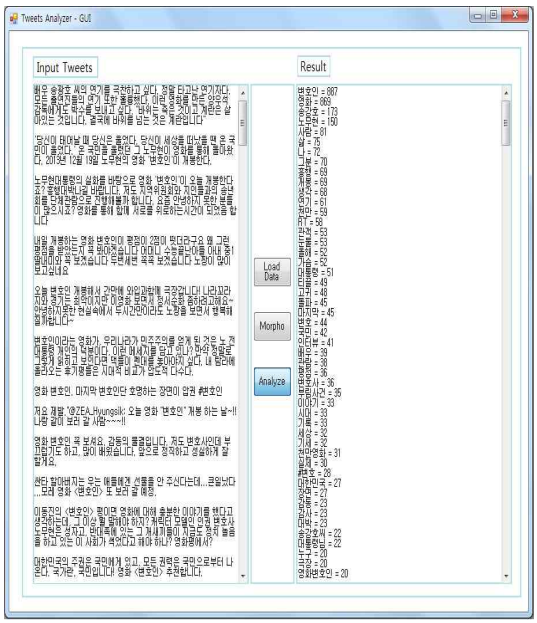

그림 8. 명사의 빈도수 결과

V. 영화 <변호인>의 흥행 요인 분석

영화 <변호인>은 수많은 논란에도 불구하고 개봉 33일 만에 천만을 돌파하며 흥행에 성공하였다. IV장의 주요 토픽 추출 결과에서도 나타났듯이 영화에 대한 관객들의 관심은 '부림사건', '노무현'이라는 영화의 소재와 '배우', '연기' '이야기', '장면', '감독' 등과 같이 영화의 속성에 집중되었다. 따라서 영화 <변호인>의 흥행에 가장 긍정적인 영향을 미친 영화적 속성을 스토리와 감독, 배우로 요약할 수 있다. 이 장에서는 스토리와 감독, 배우에 대해 언급한 트윗의 발생 비율과 이 세 가지 요소에 대한 관객들의 구체적인 반응을 살펴보고자 한다.

1. 주요 속성에 대한 트윗 발생 비율

[그림 9]는 토픽 '스토리', '배우', '감독'에 대한 트윗 발생 비율을 나타낸 것이다.

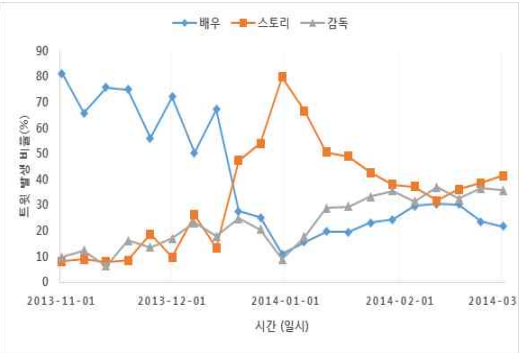

그림 9. 토픽 '스토리', '배우', '감독'의 트윗 발생 비율

'스토리' 키워드는 영화가 개봉되기 전에는 낮은 비율을 보이다가 영화가 개봉하면서 급격하게 높아진 반면 배우 '송강호'에 대한 언급은 개봉 전부터 높은 비율을 보였다. 배우 '송강호'는 이미 많은 영화를 통해 관객들에게 높은 신뢰도를 준 배우이며, 송강호가 출현했던 <설국열차>, <관상>이 2013년에 개봉되어 연이어 흥행에 성공하면서 영화 <변호인>에 대해서도 관객들이 높은 기대를 가진 것으로 보인다. <변호인>은 양우석 감독의 데뷔작으로 감독으로서의 대중적 인지도가 낮았기 때문에 '감독'에 대한 관객들의 반응은 영화가 개봉된 후에도 '스토리'와 '송강호'에 비해 낮은 비율을 보였다. 하지만 영화가 천만을 넘기면서 후반부로 갈수록 트윗에서의 발생 비율이 높게 나타났다.

2. 영화 속성에 대한 관객들의 반응

2.1 영웅 신화 구조를 차용한 감동적인 스토리

조셉 캠벨이 제시한 영웅 신화의 구조는 그동안 수많은 영화와 소설에 차용되며 감동과 재미를 제공하였다. 캠벨이 말한 영웅의 여정을 살펴보면 영웅은 시대적 소명을 깨닫고, 변모하고, 새로운 세계에 도전하여 세상을 바꾸려고 한다. 영화 <변호인>의 스토리 구조는 이러한 영웅 신화의 구조와 정확히 일치한다. 영화는 물질적 성공만을 향해 달려가는 속물 변호사 송우석이 공권력의 폐해와 부당함에 맞서는 용기 있는 변호사로 변모하는 과정을 감동적으로 그려내고 있다. [그림 10]에서 보듯이 관객들은 '좋은 영화', '감동적인 이야기', '카타르시스를 느낀다', '높은 완성도' 등과 같은 반응을 보이며 영화의 스토리에 대해 긍정적인 의견을 표현하였다.

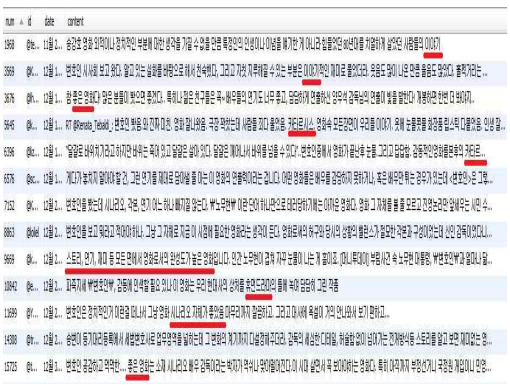

그림 10. 영화의 스토리에 대한 트위터의 반응

2.2 휴먼 드라마를 강조한 감독의 연출력

양우석 감독은 정치라는 민감한 코드를 최대한 배제하고 인권변호사로 변모해가는 주인공을 통해 휴먼드라마라는 색채를 강조하였다. 그 결과 정치적 소재에 대해 관객이 반감이 들지 않도록 하면서 현실성과 공감을 동시에 이끌어내는데 성공하였다. [그림 11]에서처

럼 감독의 연출력에 대해 관객들은 트위터에서 '양우석 감독의 발견', '세심한 디테일', '양우석 감독님의 연출이 빛을 발한다' 등과 같은 의견을 표현하며 감독의 연출력에 대해 긍정적으로 언급하였다.

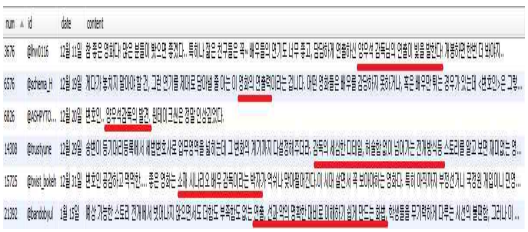

그림 11. 감독의 연출력에 대한 트위터의 반응

2.3 배우들의 열연

극 중 송우석 변호사를 연기한 송강호를 포함하여 배우들의 열연은 관객을 영화관으로 이끈 중요한 원동력이다. 관객들은 '연기의 신', '송강호의 진면목' 등의 반응을 보이는 등 트위터에서 배우들의 뛰어난 연기에 대해 찬사가 이어졌다. [그림 12]는 배우들에 대한 관개들의 의견이다.

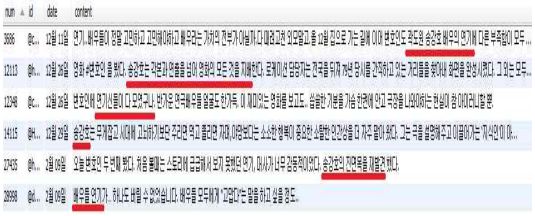

그림 12. 배우들의 연기에 대한 트위터의 반응

영화 <변호인>은 블록버스터 장르나 거대한 자본의 공세가 아닌 스토리의 힘, 감독의 연출, 그리고 신뢰할 수 있는 배우가 흥행에 더 중요한 요소라는 사실을 보여주었다.

3. 토픽 분석 결과의 유효성 검증

기존의 영화 수요 예측 연구에서 많은 연구자들은 영화 흥행에 유의미한 영향을 미치는 영화의 속성으로 스토리, 감독, 스타 배우, 제작비, 장르 등을 제시하였다. 본 연구에서 영화 <변호인>의 트윗에 대한 빈도수 기반의 토픽을 추출한 결과 '스토리'와 '배우', '감독'과 같은 영화의 속성을 나타내는 토픽 키워드들이 포함되어 있었다. 이 키워드들은 관객들이 해당 영화에 대해 가장 관심을 갖는 것이 무엇인지를 보여주는 지표이자 관객들이 영화를 선택한 구체적인 동기를 나타낸다. 본 연구는 토픽 추출에서 그치지 않고 토픽들이 시간에 따라 그 비율이 어떻게 정량적으로 변화하는지를 보여주었다.

VI. 결론 및 향후 연구

전통적인 매스미디어들은 홍보를 통해 사용자들에게 상품의 인지도를 높일 수는 있었지만 사용자들이 어떤 동기와 이유로 상품이나 대상을 선택하였는지에 대해서는 설명하지 못했다. 하지만 소셜미디어는 실시간으로 자신의 생각과 의견을 생산하고 공유할 수 있는 환경을 사용자에게 제공해 줌으로써 사용자가 특정 대상을 선호하거가 선택하는 근거가 무엇인지를 밝혀줄 수 있는 중요한 정보원으로서의 역할을 하고 있다.

소셜미디어를 통해 올라오는 영화에 대한 사용자들의 의견과 반응은 영화의 흥행 추이를 예측하고 분석하는데 매우 유용한 데이터들이다. 본 연구는 영화 <변호인>을 천만 흥행으로 이끈 요소들이 무엇인지를 밝혀내는 것을 목적으로 하였다. 이를 위해 해당 영화에 대해 언급한 트윗을 수집하고 전처리 과정을 거쳐 고빈도 명사를 추출하였다. 트위터에서 빈번하게 언급되었던 고빈도 명사들은 해당 영화에 대해 관객들의 관심을 대표하는 토픽으로 간주하였다. 그리고 토픽 분석을 통해 영화에 대해 관객들이 선호하는 대상이자 선택의 동기가 되는 요소로 스토리, 배우, 감독의 연출력임을 밝혀내고 이 세 개의 영화적 속성들이 결국 영화의 흥행에 영향을 미친 주요 요인임을 제시하였다.

SNS 데이터를 이용한 기존의 영화 흥행 연구는 주로 주기별 SNS 데이터의 변화량과 영화 흥행간의 상관성을 보여주거나 SNS 데이터에 대해 감성의 극성을 결정하는 오피니언 마이닝에 집중되어 있었다. 하지만 정량적 분석만으로는 영화의 흥행 요인을 밝히는데 한계가 있으며, 감성 극성의 결과 값만으로는 관객들이 영화의

어떤 속성을 좋아하는지는 알 수 없다. 본 연구는 트윗으로부터 빈도수 기반의 토픽을 추출하여 관객들이 관심을 갖고 있는 영화의 속성을 제시하고 그 속성들에 대해 관객들의 반응을 분석했다는 점에서 의미가 있다.

향후에는 좀 더 다양한 장르의 영화들을 대상으로 장르별 흥행 요인을 제시하는 것이 필요하다. 장르에 따른 흥행 요인 분석은 영화 산업에서 중요한 정보원으로 활용될 수 있다. 또한 영화 소비 단계뿐만 아니라 기획 및 제작 단계에서도 소셜 빅데이터를 이용하여 관객들의 선호도를 미리 예측함으로써 영화 흥행의 성공의 가능성을 높일 수 있을 것으로 기대된다.

참 고 문 헌

[1] 시로타 마코타, 김성재 역, *빅데이터의 충격*, 한빛미디어, p.26, 2013.

[2] 송길영, "Social Big Data Mining Service : 활용의 예", 한국 IT 서비스 학회 학술대회 논문집, Vol.2012, No.3, pp.161-187, 2012.

[3] 황승구, 최완, 장명길, 이미영, 허성진, *빅데이터 플랫폼 전략*, 전자신문사, p.150, 2013.

[4] 강지훈, 박찬희, 도형록, 김성범, "데이터마이닝 기법을 활용한 영화 흥행 실적 예측 기법", 대한산업공학회 춘계학술대회논문집, Vol.2014, No.5, pp.142-154, 2014.

[5] 권선주, "영화 흥행성과의 분석과 예측: 뉴스와 웹사이트 데이터 이용", 한국문화경제학회 문화경제연구, 제17권, 제1호, pp.35-56, 2014.

[6] 김연형, 홍정한, "영화흥행 결정요인과 흥행성과 예측연구", 한국통계학회 논문집, 제18권, 제6호, pp.859-869, 2011.

[7] 김진욱, "영화 마케팅의 빅데이터 활용 효과에 관한 연구", Vol.8, No.2, pp.349-356, 2014.

[8] G. Mishne and N. S. Glance, "Predicting Movie Sales from Blogger Sentiment," In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp.155-158, 2006.

[9] L. Lica and M. Tuta, "Predicting Product Performance with Social Media," nformatics in education, Vol.15, No.2, pp.46-56, 2011.

[10] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics(NAACL-HLT), Association for Computational Linguistics, pp.293-296, 2010.

[11] S. Asur and B. A. Huberman, "Predicting the future with social media," in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010, Vol.1, No.6, pp.492-499, 2010.

[12] T. Kim, J. Hong, and H. Koo, "Forecasting Box-Office Revenue by Considering Social Network Services in the Korean Market," Sains Humanika, Vol.64, No.2, 2013.

[13] 김혜원, "저예산 영화 마케팅에서의 트위터 활용 방안", Vol.11, No.1, pp.111-130, 2011.

[14] 안지혜, 민병현, "영화 마케팅 채널로서 소셜미디어의 가능성 : 다큐멘터리 영화 <땅의 여자> 트위터 마케팅 사례를 중심으로", 한국콘텐츠학회논문지, Vol.11, No.6, pp.228-241, 2011.

[15] A. Elberse and B. Anand, "The effectiveness of pre-release advertising for motion pictures: An empirical investigation using a simulated market," Information Economics and Policy, Vol.19, No.3, pp.319-343, 2007.

[16] F. Zufryden, "Linking Advertising to Box Office Performance of New Film Releases-A Marketing Planning Model," Journal of Advertising Research, Vol.36, No.4, pp.29-42, 1996.

[17] 이오준, 백영태, "협업 필터링 추천 시스템을 위한 데이터 신뢰도 기반 가중치를 이용한 하이브

리드 선호도", 한국컴퓨터정보학회지논문, Vol.19, No.5, pp.61-69, 2014.
[18] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," Communications of the ACM, Vol.40, No.3, pp.66-72, 1997.
[19] R. Mukherjee, N. Sajja, and S. Sen, "A movie recommendation system-an application of voting theory in user modeling," User Modeling and User-Adapted Interaction, Vol.13, No1, pp.5-33, 2003.
[20] 강범모, 김흥규, "명사 빈도의 변화, 사회적 관심의 트렌드 : 물결 21 코퍼스[2000-2009]", 언어학, Vol.61, pp.3-38, 2011.
[21] 강성진, *군집화 기법과 문서 순위를 이용한 한국어 트윗 상의 토픽 추출*, 서울대학교 대학원 석사학위 논문, p.14, 2013.
[22] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting movie prices through dynamic social network analysis," Pcocedia-Social and Behavioral Sciences, Vol.2, No.4, pp.6423-6433, 2010.
[23] S. Kim, S. Jeon, J. Kim, Y. H. Park, and H. Yu, "Finding core topics : Topic extraction with clustering on tweet," In Cloud and Green Computing, 2012 Second International Conference, IEEE, pp.777-782, 2012.
[24] X. Ni, X.Quan, Z. Lu, L. Wenyin, and B. Hua, "Short text clustering by finding core terms," Knowledge and information systems, Vol.27, No.3, pp.345-365, 2011.
[25] 허민회, 강필성, 조성준, "오피니언 마이닝을 이용한 영화 흥행의 예측", 2013 한국경영과학회/대한산업공학회 춘계공동학술대회 논문집, Vol.2013, No.5, pp.487-500, 2013.
[26] 박선영, "SNS를 통한 구전 효과가 영화 흥행에 미치는 영향-<써니>의 사례를 중심으로", 한국콘텐츠학회논문지, Vol.12, No7, pp.40-53, 2012.
[27] http://fizziolo.gy/products/
[28] http://www.mezzomedia.co.kr/mezzomedia-social-analysis/
[29] http://wstarnews.hankyung.com/apps/news?popup=0&nid=01&c1=01&c2=01&c3=00&nkey=2013050916323 21&mode=sub_view
[30] www.pulsek.com
[31] http://ko.wikipedia.org/wiki/%EC%9B%B9_%ED%81%AC%EB%A1%A4%EB%9F%AC
[32] http://kldp.net/projects/hannanum
[33] http://docs.oracle.com/javase/7/docs/api/java/util/TreeMap.html

저 자 소 개

이 오 준(O-Joun Lee) 준회원

- 2011년 3월 ~ 현재 : 단국대학교 소프트웨어학과 학부 재학 중

<관심분야> : 추천시스템, 기계학습, 정보검색, 적응형시스템

박 승 보(Seung-Bo Park) 정회원

- 1995년 2월 : 인하대학교 전기공학과(공학사)
- 1997년 2월 : 인하대학교 전기공학과(공학석사)
- 2011년 3월 : 인하대학교 정보공학과(공학박사)
- 2013년 8월 ~ 2014년 8월 : 단국대학교 미디어콘텐츠연구원 연구원
- 2014년 10월 ~ 현재 : 인하대학교

<관심분야> : 멀티미디어 정보검색, 스토리텔 링, 감정 검출

정 다 울(Daul Chung) 준회원

▪ 2008년 3월 ~ 현재 : 단국대학교 소프트웨어학과 학부 재학 중

<관심분야> : 빅데이터, 데이터마이닝, 오피니언 마이닝

유 은 순(Eun-Soon You) 정회원

▪ 1995년 2월 : 인하대학교 불어불문학과(문학사)
▪ 2000년 10월 : 프랑스 브장송대학교 언어학(석사)
▪ 2007년 7월 : 프랑스 브장송 대학교 언어학(박사)
▪ 2012년 ~ 현재 : 단국대학교 미디어콘텐츠연구원 리서치 펠로우

<관심분야> : 빅데이터, 데이터마이닝, 스토리텔링