

# 의미연결망 분석을 활용한 영화 리뷰 시각화

김슬기\* · 김장현\*\*

성균관대학교

## A Visualization of Movie Review based on a Semantic Network Analysis

Seul-gi Kim\* · Jang Hyun Kim\*\*

Sungkyunkwan University

E-mail : april13@skku.edu / alohakim@skku.edu

### 요 약

본 연구에서는 <네이버 영화> 페이지의 리뷰 데이터를 수집하여, 출현 빈도가 높은 단어를 중심으로 영화 관람객의 반응을 시각화하는 작업을 수행하였다. 이를 위해 총 6편의 영화를 선정하여 데이터 수집 및 정제과정을 거쳤으며, 의미연결망 분석(Semantic network analysis)을 활용하여 단어 간 관계성을 파악하고자 하였다. 데이터 시각화 작업에는 UCINET과 함께 패키지가화된 NetDraw가 사용되었다. 본 연구의 시사점은 문장으로 작성된 영화 관람객의 리뷰를 키워드 중심으로 시각화하여, 소비자들의 반응을 한 눈에 확인하는 리뷰 인터페이스 구현이 가능한지 탐색하였다는 점이다.

### ABSTRACT

The aim of current research is to suggest a interface for movie reviews at a glance through semantic network analysis. The implication of this study is to systematically investigate the structure of eWoM. Specifically, by visualizing semantic networks of movie reviews this study attempts to provide a prototype of a possible review system that can check the response of movie viewer at a glance.

### 키워드

Big data, Semantic network analysis, Movie, Review

## 1. 서 론

현재 한국에서 영화 관람은 일상 속 문화생활로 자리 잡혀 있다. 이를 반증하듯, 한국 영화산업 시장의 매출은 2014년 이후 2조원 대를 유지하고 있으며, 연간 평균 관람 횟수는 1인당 4.2회로 세계 최고의 수준이다[1]. 이처럼 한국 영화산업은 전 세계적으로 주목받을만한 시장 규모를 가지고 있으며, 한국 영화 관람객을 끌어들이기 위한 다양한 행사와 마케팅이 이루어지고 있다.

영화는 경험을 소비하는 제품을 일컫는 경험재로서, 직접 관람하기 이전에는 그 품질을 판단하기가 어렵기 때문에 예고영상 또는 관람객의 추천 여부에 따라 관람 의도가 결정되는 경우가 많다. 이러한 이유로 영화 흥행은 개봉 초기에는 영화평론가의 평가, 후기에는 구전(Word of Mouth : WOM)의 영향을 많이 받는 특성이 있으며[2], 온라인

인상에 존재하는 수많은 리뷰는 관람객의 영화 관람 결정에 큰 역할을 담당하고 있다. 따라서 소비자들은 이미 다양한 경로를 통해 영화 평점 및 리뷰를 확인하고 있으며, 수많은 웹사이트와 영화 관련 플랫폼들은 각자의 인터페이스와 접근방식을 보유하고 있다.

본 연구는 온라인상에 존재하는 영화 리뷰의 키워드를 추출하고, 단어 간 관계를 시각화하는 작업을 수행하였다. 구체적으로 <네이버 영화> 페이지의 데이터를 크롤링하여, 출현 빈도가 높은 단어를 중심으로 네트워크 분석을 실시하였다. <네이버 영화> 사용자는 1에서 10 사이의 평점을 부여한 후, 140자 내외의 감상평을 작성할 수 있다. 이러한 제한된 글자 수로 인해, 리뷰 작성자는 본인이 느낀 감정이나 영화에 대한 견해를 보다 핵심적인 형태로 표현하고 있었다.

본 연구의 시사점은 문장으로 작성된 영화 관람객의 리뷰를 키워드 중심으로 시각화하여, 소비자들의 반응을 한 눈에 확인할 수 있는 리뷰 인터페이스 구현이 가능한지 탐색하였다는 점이다. 기존

\* speaker

\*\* corresponding author

연구에서는 영화 리뷰의 박스오피스 예측 가능성 등 기능적 측면에 초점을 둔 반면, 리뷰 자체를 체계적으로 시각화한 경우는 거의 없었다. 본 연구는 이러한 부분에 기여하는 바가 크다고 하겠다.

## II. 관련연구

영화 산업계에서 빅데이터를 활용한 연구는 SNS(Social Network Service)에 올라오는 영화 관련 데이터를 이용하여 영화의 흥행을 예측하는 연구, 영화 마케팅에서의 소셜미디어 활용에 대한 연구, 그리고 SNS 데이터에 내재된 개인 성향 분석을 통한 맞춤형 영화 추천 시스템 개발 연구 등으로 정리할 수 있다[3].

온라인상의 영화 리뷰를 이용한 연구 사례로는 소셜 네트워크 분석과 감정 분석을 결합한 웹 마이닝 기술을 통해 IMDB(인터넷 영화 데이터베이스)의 영화 리뷰를 분석하고, 영화의 흥행 여부를 예측한 Doshi et al.(2010)의 연구가 있다[4]. Kim et al.은 Core-Topic-based Clustering을 이용하여 특정 인기 TV 드라마의 트윗을 분석하고 의미 있는 토픽들을 추출하였고[5], Ni et al.은 트위터에 존재하는 core word를 추출하여 사용자들에 의해 많이 언급된 영화 내용이 무엇인지를 제시하고 토픽에 따라 유사한 트윗을 클러스터링 하였다[6].

다음의 선행 연구들을 통해 영화 리뷰를 이용한 흥행 예측은 의미 있는 결과를 보여주고 있으나 설명력의 절대량 자체는 약한 편이다. 또한, 영화 마케팅에서의 소셜미디어 활용에 관한 연구는 주로 설득 이론적 관점에서의 접근이 많았고, 추천 시스템은 속성 분석을 통한 관계성 도출에는 의미가 있으나 영화 리뷰의 자체에 대한 분석은 미약한 편이다.

본 연구에서는 영화 리뷰의 핵심단어를 추출하고, 유사한 단어들 간의 군집화를 통해 각 영화 리뷰를 구성하는 주된 이슈가 무엇인지 시각적으로 제시하고자 하였다. 이와 같이 리뷰 내용 자체의 특성을 분석하고, 키워드를 시각화하였다는 것이 본 연구의 시사점이라고 하겠다.

## III. 연구방법



그림 1. 연구절차

본 연구에서는 <네이버 영화>의 리뷰를 활용하였고, 구체적인 연구절차는 그림 1과 같다. 먼저, 2016년 흥행작 3편과 네티즌 평점 기준 7점미만의 영화 3편을 선정하여 영화 리뷰를 수집하였다.

표 1. 영화별 정보 및 집계 데이터 (2018.8.22.기준)

영화제목	장르	개봉년도	리뷰수	네티즌 평점
부산행	스릴러	2016	57,935	8.00
터널	드라마	2016	24,522	8.37
밀정	액션	2016	23,234	8.42
엽기적인 그녀2	멜로/로맨스	2016	1,870	2.84
리얼	액션	2017	12,071	4.25
희생부활자	미스터리	2017	2,804	6.14

크롤링된 데이터의 형태소 분리에는 한국어 형태소 분석 라이브러리인 KoNLPy(Korean natural language processing in Python)가 사용되었다. 이를 통해 단어는 출현 빈도순으로 정렬되고, 유의미한 단어들끼리 묶여 분석이 가능한 상태로 가공되었다. 데이터는 Microsoft Excel 형태로 추출되었고, 아래의 기준으로 데이터 정제 작업을 실시하였다.

- 1) 관사, 조사와 같은 불용어 삭제
- 2) 같은 의미의 단어 통합 (예: 배우이름(차태현-태현), 연기-연기력)
- 3) 불필요한 단어 삭제 (예: 영화제목, 그냥, 감탄사 등)

다음의 정제작업을 거친 데이터 중에서 출현 빈도를 기준으로 각 영화당 상위 100개의 단어만이 연구에 사용되었다. 또한, 단어들 간의 연결성 및 중심성을 파악하기 위해 UCINET을, 키워드를 시각화하기 위해 NetDraw를 사용하였다.

본 연구에서는 단어 간 네트워크 분석에 활용되는 여러 지표 중 중심성(Centrality), 그중에서도 위세 중심성(Eigenvector centrality)을 활용하였다. 위세 중심성을 선택한 이유는, 네트워크 시각화에 활용되는 단어 중에서 어떤 노드가 가장 영향력이 있는지를 시각적으로 나타내기 위해서이다. 위세 중심성은 연결된 이웃 노드의 중심성이 높으면, 그 노드의 중심성도 함께 증가한다는 관점을 반영한다. 즉, 연결된 노드들의 중심성에 가중치를 부여하여 측정하기 때문에 영향력이 낮은 여러 노드와 많은 연결이 있는 경우보다 중요한 몇 개 노드에 연결되어 있는 경우가 해당 노드의 중심성이 높다고 볼 수 있다[7]. 네트워크 분석 및 시각화 단계 전에 이미 상위 빈도수를 기준으로 단어를 정리했지만, 어떤 단어가 전체 리뷰에서 큰 영향력을 차지하고 있는지는 알 수 없기 때문에 위세 중심성이 높은 노드가 눈에 띄도록 하였다.

데이터 시각화 작업에는 UCINET과 함께 패키지화된 NetDraw가 사용되었고, 단어 간 유사성을 기준으로 분석되는 CONCOR(Convergence of iterated CORrelations) 기법을 활용하였다. CONCOR 분석을 통해 유사성을 중심으로 단어 군집이 형성되었고, 같은 군집에 포함된 다른 단어들을 통해 해당 단어의 성격을 유추할 수 있도록 하였다. 즉, 단순히 빈도수가 높은 단어들의 나열만으로는 리뷰 작



