JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 23, No. 1: 1~6, Jan. 2019

# 의미연결망 분석을 활용한 영화 리뷰 시각화

김슬기[1] · 김장현[2*]

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

**Seulgi Kim[1] · Jang Hyun Kim[2*]**

[1]Graduate Student, Department of Interaction Science, Sungkyunkwan University, Seoul, 03063 Korea
[2*]Associate Professor, Department of Interaction Science, Sungkyunkwan University, Seoul, 03063 Korea

## 요 약

본 연구는 <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 패키지화된 NetDraw가 사용되었다. 본 연구의 시사점은 문장으로 작성된 영화 관람객의 리뷰를 키워드 중심으로 시각화하여, 소비자들의 반응을 한 눈에 확인하는 리뷰 인터페이스 구현이 가능한지 탐색하였다는 점이다. 본 연구를 통해 영화 리뷰를 구성하는 키워드를 시각화하고, 리뷰 내용에서 영화별 특성의 차이를 확인하였다는 점에서 본 연구가 의미를 가진다고 하겠다. 후속 연구는 보다 많은 영화의 리뷰를 활용할 필요성이 제기되며, 각 영화별 리뷰의 수도 비슷한 양으로 맞추어 연구에 활용해야 할 것이다.

## ABSTRACT

This study visualized users reaction about movies based on keywords with high frequency. For this work, we collected data of movie reviews on <Naver Movie>. A total of six movies were selected, and we conducted the work of data gathering and preprocessing. Semantic network analysis was used to understand the relationship among keywords. Also, NetDraw, packaged with UCINET, was used for data visualization. In this study, we identified the differences in characteristics of review contents regarding each movie. The implication of this study is that we visualized movie reviews made by sentence as keywords and explored whether it is possible to construct the interface to check users' reaction at a glance. We suggest that further studies use more diverse movie reviews, and the number of reviews for each movie is used in similar quantities for research.

**키워드** : 빅데이터, 의미연결망 분석, 영화, 온라인 리뷰

Keywords : Big data, Semantic network analysis, Movie, Online review

**Received** 9 September 2018, **Revised** 27 September 2018, **Accepted** 7 November 2018

*** Corresponding Author** Jang Hyun Kim(E-mail:alohakim@skku.edu, Tel:+82-2-740-1868)
Associate Professor, Department of Interaction Science, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://doi.org/10.6109/jkiice.2019.23.1.1 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

## Ⅰ. 서 론

소셜미디어의 발달을 통해 소비자 간의 지식 및 경험의 공유가 가능해졌고, 이러한 과정에서 생산되는 콘텐츠는 웹상에서 방대한 양의 데이터로서 그 역할을 다하고 있다[1]. 이러한 온라인 토론과 관련하여 자주 언급되는 주제 중 하나는 영화인데, 평론가뿐만 아니라 영화 관람객으로부터 작성되는 리뷰의 영향력이 갈수록 높아지면서, 영화 리뷰를 주고받을 수 있는 웹사이트 또는 모바일 플랫폼은 영화 관계자 및 소비자들로부터 정보가 끊임없이 생산되는 원천이 되었다[2].

영화는 경험을 소비하는 제품을 일컫는 경험재로서, 직접 관람하기 이전에는 그 품질을 판단하기가 어렵기 때문에 예고영상 또는 관람객의 추천 여부에 따라 관람의도가 결정되는 경우가 많다. 실제로 업계 전문가들은 온라인 구전이 영화의 저력을 결정짓는 중요한 결정요인이라고 말하며, 궁극적으로 영화의 흥행 성공 또는 실패를 이끌 수 있다고 설명한다[3]. 이렇듯 온라인상에 존재하는 수많은 리뷰는 관람객의 영화 관람 결정에 큰 역할을 담당하고 있으며, 소비자들은 이미 다양한 경로를 통해 영화 평점 및 리뷰에 접근하고 있다. 이러한 흐름에 따라 수많은 웹사이트와 영화 관련 플랫폼이 등장했고, 다른 플랫폼과 차별화를 통한 더 많은 사용자의 유입을 도모하기 위해 그들은 각자의 인터페이스와 접근 방식을 적용하고 있다. 그만큼 리뷰를 제시하는 방식이나 디자인 등의 역할이 나날이 중요해지고 있는 상황이다.

본 연구는 온라인상에 존재하는 영화 리뷰의 수집하여, 출현빈도를 기준으로 키워드를 추출하였고, 단어 간 관계성을 네트워크 분석을 통해 시각화하는 작업을 수행하였다. 구체적으로 <네이버 영화> 페이지의 리뷰 데이터를 크롤링하여, 의미 연결망 분석을 실시하였다. <네이버 영화> 사용자는 1에서 10 사이의 평점을 부여한 후, 140자 내외의 감상평을 작성할 수 있다. 작성 가능한 글자 수가 제한되어 있기 때문에, 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. 대부분의 영화 리뷰 사이트 혹은 플랫폼들은 문장형의 리뷰 제시 인터페이스를 활용하고 있으며, 관련 선행 연구들은 영화 리뷰의 박스오피스 예측 가능성과 리뷰의 효과성을 검증하는 등 기능적 측면에 초점을 둔 반면, 리뷰 자체를 체계적으로 시각화한 경우는 없었던 것으로 파악된다. 따라서 본 연구를 통해 영화에 대한 반응을 보다 자세히 확인하였다는 점에서 기여하는 바가 있다고 하겠다.

## Ⅱ. 관련 연구

빅데이터 분석 기술이 발달하면서 데이터 마이닝, 오피니언 마이닝, 텍스트 마이닝, 웹 마이닝, 소셜 마이닝 등 다양한 기법을 통해 빅데이터 관련 연구가 수행되어 왔다[4]. 이러한 흐름에 발 맞춰 영화 산업계에서도 빅데이터를 활용화여 다양한 연구를 진행해왔다. 특정 영화의 트위터 내용을 분석하여 관객들이 영화의 어떤 속성을 선호하는지 살펴본 연구[5], 온라인 리뷰의 길이와 영화 흥행의 관계성을 검증한 연구[6], 위키피디아의 편집자와 독자의 활동 수준에 따른 초기 영화 흥행여부를 예측한 연구[7] 등 다양한 데이터를 활용한 연구들이 수행되어 왔다.

영화평과 평점을 이용해 감성 분석을 실시한 오연주, 채수환 (2015)의 연구는 온라인 사용자가 선택한 평점에 비해 감성 문장 집합을 적용하여 측정한 영화평의 감성 점수가 사용자들의 의견을 더 잘 반영한다는 것을 확인하였다[8]. 또한, Doshi 외 (2010)는 소셜 네트워크 분석과 감성 분석을 결합하는 방법이 트렌트를 예측하는데 얼마나 효과적인지를 검증하였다. 구체적으로, 그들의 연구는 IMDB와 Rotten Tomatoes, 박스오피스 정보 제공 웹사이트에서 데이터를 수집하여, 개봉 초기 영화들의 흥행 여부를 예측하였다 [9].

빅데이터에서 키워드를 추출하는 방법을 활용한 연구로는 Kim 외 (2012)가 있다. 저자들은 Core-Topic-based Clustering (CTC)라고 불리는 기법을 제안하면서, 여러 인기 TV 드라마의 트윗을 추출하고, 유의미한 토픽을 추출하였는데, 기존의 알고리즘보다 클러스터링 작업이 더 효과적이라고 주장하였다 [10].

이 밖에도 많은 선행 연구에서 영화 리뷰를 활용한 데이터마이닝 기법을 통해 영화의 흥행을 예측하고, 다양한 마케팅 전략을 제안해왔다. 이러한 연구를 통해 영화

의 흥행을 예측하는 의미 있는 결과를 발견하고, 영화 추천 시스템을 개발하는 등의 연구적 성과는 있으나, 영화 리뷰의 내용적 측면에 집중하여 분석한 연구는 많지 않은 편이다.

따라서 본 연구는 영화 리뷰에서 핵심 단어를 추출하고, 유사한 단어들 간의 군집화를 통해 각 영화 리뷰를 구성하는 주된 이슈가 무엇인지 시각적으로 제시하고자 하였다. 이를 통해 영화 리뷰를 구성하는 키워드를 시각화하고, 리뷰 내용에서 영화별 특성의 차이를 확인하였다는 점에서 본 연구가 의미를 가진다고 하겠다.

## Ⅲ. 연구 방법

본 연구에서는 <네이버 영화>의 리뷰를 수집하여 분석에 활용하였다. 먼저, 2016년 흥행작 3편과 네티즌 평점 기준 7점미만의 영화 3편을 선정하여 영화 리뷰를 수집하였다. 영화 선정의 기준을 다음과 같이 설정한 이유는 평이 좋은 영화와 그렇지 않은 영화 간의 차이를 비교하기 위해서이다. 표 1은 영화에 대한 간략한 정보와 리뷰의 수를 정리한 것이다.

**Table. 1** Movie Information and Aggregate Data

| Title | Genre | Release Year | The number of reviews | Users' rating |
|---|---|---|---|---|
| Train to Busan | Thriller | 2016 | 57,935 | 8.00 |
| Tunnel | Drama | 2016 | 24,522 | 8.37 |
| The Age of Shadows | Action | 2016 | 23,234 | 8.42 |
| My New Sassy Girl | Melodrama / Romance | 2016 | 1,870 | 2.84 |
| REAL | Action | 2017 | 12,071 | 4.25 |
| RV: Resurrected Victims | Mystery | 2017 | 2,804 | 6.14 |

By creation date: 2018.8.22.

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 노드의 중심성이 높으면, 그 노드와 연결된 다른 노드의 중심성도 함께 높아진다는 관점을 반영한다. 즉, 연결된 노드들의 중심성에 따라 가중치를 부여하기 때문에, 영향력이 높은 주변 노드에 연결되어 있는 경우의 중심성은 낮은 영향력을 가진 노드들과 많이 연결된 경우보다 더 높다고 할 수 있다[11]. 본 연구에서는 네트워크 분석 및 시각화 단계 전에 이미 상위 빈도수를 기준으로 단어를 정리했다. 그러나 이것만으로는 어떤 단어가 전체 리뷰에서 상대적으로 더 큰 영향력을 발휘하고 있는지 알 수 없기 때문에 위세 중심성이 높은 노드가 눈에 띄도록 설정하였다.

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 작성자가 영화를 추천하기 위해 사용한 것인지, 혹은 그 반대인지를 알 수 없기 때문에, 단어가 포함된 클러스터링 그룹을 통해 해당 단어의 사용 의도를 추측할 수 있도록 CONCOR 분석을 활용하였다.

## Ⅳ. 분석 결과

전처리를 통해 의미 있는 형태소를 분리해 내고, 네트워크 분석을 통해 각 영화별로 8개의 클러스터링 그룹과 100개의 단어를 도출하였다. 또한, 영향력 있는 노드가 눈에 잘 띄도록 위세 중심성에 따라 텍스트의 크기가 달라지도록 설정하였다. 마지막으로 CONCOR 분석을 통해 네트워크 내에서 단어 간 연결 관계 및 패턴이 한눈에 파악될 수 있도록 시각화하였다. 유사성을 바탕으로 형성된 군집은 특정 단어가 어떤 의도로 사용되었는지 유추 가능하도록 하며, 리뷰 전체를 아우르는 공통된 주제도 파악할 수 있게 되었다.

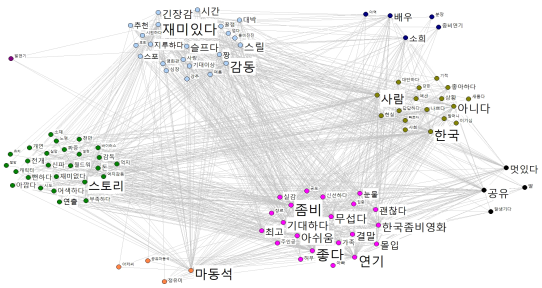


**Fig. 1** Train to Busan

먼저, 그림 1은 영화 <부산행>의 네트워크 관계를 이미지화 한 것이다. <부산행>은 2016년 최고 흥행작으로, 월등하게 많은 리뷰가 작성되었다. 위세 중심성이 높은 노드는 주로 감정표현과 '좀비', '연기'와 연결된다는 것을 알 수 있으며, '스토리' 함께 군집을 이루는 부정적 노드가 많은 것으로 보아 스토리에 다소 아쉬운 평이 많다는 추측할 수 있다.

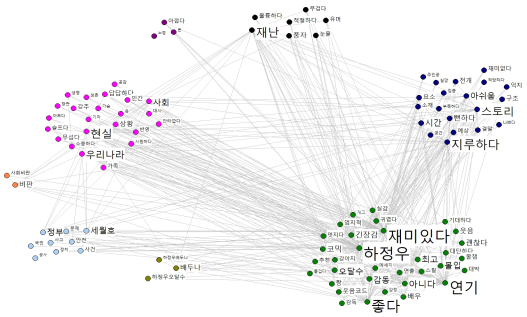


**Fig. 2** Tunnel

그림 2는 영화 <터널>와 관련된 데이터를 시각화한 것으로, '하정우'라는 배우의 이름과 '재미있다'가 가장 영향력 있는 노드라는 것을 확인할 수 있다. 또한, '스토리'에 대한 아쉬움과 영화의 소재인 재난을 둘러싼 현실 상황과 관련된 노드들이 많은 것을 확인할 수 있다.

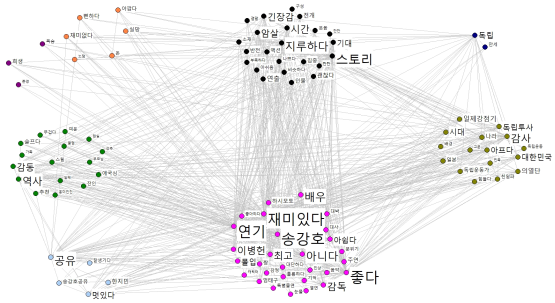


**Fig. 3** The Age of Shadows

영화 <밀정>의 키워드 간 관계는 그림 3을 통해 확인할 수 있는데, 배우와 연기가 주된 감상평이라는 것을 알 수 있다. 뿐만 아니라, 이 영화 또한 스토리에 대한 평이 많았으며, 부정적 노드가 많지 않을 것을 보아 평점을 보지 않아도 전반적인 리뷰의 톤을 추측할 수 있다.

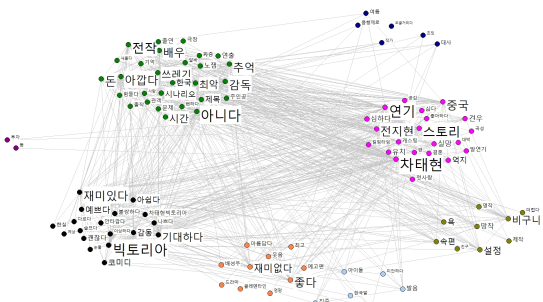


**Fig. 4** My New Sassy Girl

그림 4는 영화 <엽기적인 그녀2>의 데이터를 분석하고 시각화 한 결과이다. 본 영화는 부정적 노드의 크기가 이전의 3편에 비해 많은 것을 확인할 수 있다. 또한, 이전 3편의 영화에서는 나타나지 않았던 다소 과격한 표현들도 등장하고 있는 것을 보아 영화에 대한 관람객의 반응이 부정적인 경우가 많음을 알 수 있다.

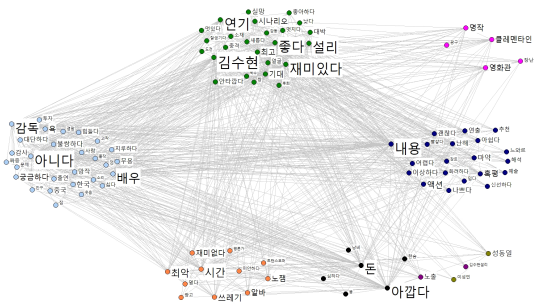


**Fig. 5** REAL

그림 5는 영화 <리얼>의 리뷰를 네트워크 분석한 결과이다. 다음의 이미지를 통해, 같은 군집에 포함된 '돈'과 '아깝다'가 영향력이 큰 노드로 확인되었다. <엽기적인 그녀2>와 마찬가지로 다소 과격한 표현이 다수 포함되어, 낮은 평점과 리뷰 내용이 일맥상통한다고 볼 수 있다.

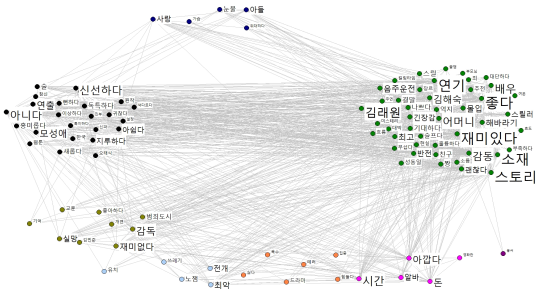


**Fig. 6** RV: Resurrected Victims

마지막으로 그림 6은 영화 <희생부활자>의 네트워크 분석 결과이다. <희생부활자>는 배우와 연기, 스토리와 관련된 노드들이 하나의 군집을 이루고 있어, 이와 관련된 언급이 많은 것을 알 수 있다. 하지만 부정적 노드들이 넓은 범위에 분포되어 있다.

## Ⅴ. 결론 및 시사점

본 연구의 목표는 의미연결망 분석을 통해 영화 리뷰를 구성하는 키워드를 시각화함으로써, 영화 리뷰를 한 눈에 확인하는 것이었다. 이를 위해 영향력이 큰 노드가 한 눈에 확인될 수 있도록 텍스트의 크기를 조절하였고, 리뷰를 구성하는 주된 토픽을 8개의 그룹으로 클러스터링하였다. 이를 통해 함께 군집을 이룬 노드가 주변 단어의 사용 의도를 파악하는 데 도움을 줄 것으로 예측하였다.

6편의 영화를 분석한 결과, 장르와 평판에 상관없이 영화의 주요 구성요소인 "감독, 배우, 스토리"와 "재미있다, 재미없다, 기대, 감동" 등의 영화 감상과 관련된 단어들이 자주 등장하는 것을 확인할 수 있었다. 반면에 온라인상에서 평판이 좋지 않은 영화에만 공통적으로 등장하는 단어들도 다수 발견되었고, 이러한 단어들의 위세 중심성 또한 높은 것을 확인할 수 있었다. (예: 알바, 쓰레기, 클레멘타인 등) 따라서 키워드를 통해 완성된 네트워크가 특정 영화의 전반적인 평판을 확인하는 데 도움을 주었다. 반면에 평판이 좋은 영화의 네트워크는 클러스터링 그룹을 이루는 노드 간의 의미가 자연스럽게 연결되어, 전반적인 영화평의 톤을 읽어내는 데 도움을 주었다.

CONCOR 분석을 활용한 네트워크 이미지를 통해서는 영화의 특성을 확인할 수 있었다. 예를 들어 배우와 관련된 군집의 크기가 크다면, 배우 또는 주요 배역이 크게 부각되는 영화라는 추측이 가능해진다. 마찬가지로 영화 감상평과 관련된 군집이 크다면, 관람객들이 전반적인 영화 분위기와 관련된 감상평을 주로 언급한다는 것을 알 수 있다.

뿐만 아니라 같은 군집에 속한 다른 단어들과의 비교를 통해, 특정 단어의 사용 의도를 파악할 수 있다. 예를 들어 긍정의 뜻을 가진 단어라 할지라도 문장의 전체 맥락이 클러스터 상에 드러나지 않는다면, 리뷰 작성자가 추천 의도로 그 단어를 사용했다고 단정하기 어렵다. 그렇기 때문에 함께 묶인 다른 단어들의 역할이 중요해진다. 따라서 단어 클릭 시 해당 단어가 사용된 전체 리뷰를 자동 분류하여 이용자들에게 제시하는 인터페이스를 구현한다면, 키워드만 제시했을 때의 단점을 보완함과 동시에 사용자들에게 리뷰를 보다 효과적으로 제시할 수 있을 것으로 기대된다.

본 연구는 방대한 양의 리뷰 속에서 주목할 만한 키워드를 찾아내어, 영화 관람객들의 반응을 한 번에 확인할 수 있는 방안을 모색하고자 했다. 본 연구의 한계로는 특정한 몇 편의 영화로만 연구를 진행했다는 점과 각 영화별로 리뷰의 수가 차이가 난다는 점이다. 본 연구는 영화 리뷰의 영향력 및 리뷰 인터페이스 구축을 위한 탐

색적 연구로서, 후속 연구에서는 이러한 한계를 보완해야할 것으로 보인다. 뿐만 아니라, 보다 많은 영화의 리뷰를 연구에 활용할 필요성이 제기되며, 해외 영화 전문 사이트로의 연구모델 확장을 통해 나라별 반응 및 키워드의 차이를 비교하는 연구도 필요하다고 본다.


## ACKNOWLEDGEMENT

This research was conducted as a result of Software-Centered University Support Project from Ministry of Science and ICT / Institute for Information and Communication Technology Promotion (2015-0-00914)


## REFERENCES


[ 1 ] J. H. Kim, and K. R. Bhatele, "Recognition using Cyber bullying in view of Semantic-Enhanced Minimized Auto-Encoder," *Asia-pacific Journal of Convergent Research Interchange, HSST*, vol. 2, no. 4, pp. 7-14, Dec. 2016.

[ 2 ] J. A. Yeap, J. Ignatius, and T. Ramayah, "Determining consumers' most preferred eWOM platform for movie reviews: A fuzzy analytic hierarchy process approach," *Computers in Human Behavior*, vol. 31, pp. 250-258, 2014.

[ 3 ] A. Elberse, and J. Eliashberg, "Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures," *Marketing Science*, vol. 22, no. 3, pp. 329-354, 2003.

[ 4 ] C. H. Ban, and D. H. Kim, "Analysis of University Department Name using the R," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no. 6, pp. 829-834, 2018.

[ 5 ] O. Lee, S. B. Park, D. Chung, and E. S. You, "Movie box-office analysis using social big data," *The Journal of the Korea Contents Association*, vol. 14, no. 10, pp. 527-538, 2014.

[ 6 ] Y. H. Cho, Y. S. Park, and H. J. Kim, "Changes in Review Length Based on the Popularity of Movies Using Big Data," *The Journal of the Korea Contents Association*, vol. 18, no. 5, pp. 367-375, 2018.

[ 7 ] M. Mestyán, T. Yasseri, and J. Kertész, "Early prediction of movie box office success based on Wikipedia activity big data," *PloS one*, vol. 8, no. 8, 2013, e71226.

[ 8 ] Y. J. Oh, and S. H. Chae, "Movie Rating Inference by Construction of Movie Sentiment Sentence using Movie comments and ratings," *Journal of Internet Computing and Services*, vo. 16, no. 2, pp. 41-48, 2015.

[ 9 ] L. Doshi, J. Krauss, S. Nann, and P. Gloor, "Predicting movie prices through dynamic social network analysis," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 4, pp. 6423-6433, 2010.

[10] 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 (CGC), 2012 Second International Conference IEEE*, pp. 777-782, Nov. 2012.

[11] J. Y. Kim, and S. H. Lee, "A study on the collaboration network analysis of document delivery service in science and technology," *Journal of Korean Library and Information Science Society*, vol. 44, no. 4, pp. 443-463, 2013.



**김슬기(Seulgi Kim)**

인터랙션사이언스학과 석사과정

※관심분야 : 소셜/시멘틱 네트워크 분석, 소셜미디어, HCI

**김장현(Jang Hyun Kim)**

커뮤니케이션학 박사

※관심분야 : 데이터사이언스, 소셜/시멘틱 네트워크 분석, 소셜미디어