TF-IDF 기반 콘텐츠 평점 분석 웹 서비스

최보영 ○ 박상근
경희대학교
qhdud0330@khu.ac.kr, sk.park@khu.ac.kr

A Web Service for TF-IDF-Based Rating Analysis of Review Contents

Bo-Young Choi○ Sang-Keun Park
Department of Software Convergence, Kyunghee University

요 약

본 연구에서는 콘텐츠 시청자가 작성한 영화 리뷰와 점수를 통해 평점 분석을 수행했다. TF-IDF를 활용하여 리뷰 속에서 의미 있는 키워드를 추출하고, 키워드가 포함된 리뷰의 점수를 통해 키워드별 점수를 산출했다. 평점을 키워드별 점수를 통해 세분화하여 방사형 그래프로 제공하고 키워드 버튼을 통해 키워드가 포함된 실제 리뷰를 제공함으로써 콘텐츠와 사용자 리뷰를 빠르게 파악할 수 있는 서비스를 제공한다.

1. 서 론

Netflix, Disney+, TVING, 쿠팡플레이 등 전 세계적으로 OTT 시장이 발전하고 있다. OTT 시장 점유율이 가장 높은 Netflix의 실적 보고서[1]에 따르면 2023년 2분기까지의 총 가입자 수가 2억 3839만 명이라고 한다. 또한 Netflix에서는 올해 총 34편의 한국 콘텐츠를 올릴 예정으로 OTT가 제작하는 콘텐츠가 많아졌다. 이렇게 많은 콘텐츠 속에서 본인 취향에 맞는 콘텐츠 한 편을 찾아보기 위해 콘텐츠별 리뷰, 평점을 제공하는 서비스를 참고하는 사람이 많다.

IMDB에서는 총 평점, 회차별 평점, 나라별 평점을 제공하고, Rotten Tomatoes에서는 평론가들의 긍정적 평가 비율과 3.5점 이상의 점수를 준 사용자의 비율을 제공한다. Daum 영화 서비스는 네티즌 평점을 제공해서 사람들의 전반적인 반응을 확인할 수 있다. 하지만 대부분의 콘텐츠 리뷰, 평점 서비스는 각 콘텐츠에 대한 평가에 어떤 기준이 적용되었는지 명확하지 않아서 시청할 콘텐츠를 선택할 때 어려움이 있다.

보다 풍부한 콘텐츠 정보를 제공하기 위해, NAVER 영화 서비스는 사용자 리뷰에 OST, 스토리, 연출, 영상미, 인기라는 5가지의 감상 포인트 비율을 추가로 제공해 보다 풍부한 평점 정보를 제공한다. 하지만 고정된 5가지 감상 포인트가 모든 콘텐츠에 똑같이 적용되고, 각 항목이 평점에 얼마나 영향을 주었는지 알 수 없다는 한계가 존재한다. 해당 감상 포인트의 긍부정 비율과 감상 포인트별 리뷰를 분석한 연구[2]가 수행되었지만, 여전히 모든 작품에 대해 똑같은 기준의 감상 포인트를 제공한다는 한계가 있었다. 각 콘텐츠에 특화된 리뷰 정화 추출을 위해, 의미연결망 분석으로 영화 리뷰를 구성하는 키워드를 추출하고, 키워드를 영향력을 기준으로 시각화한 연구[3]가 있지만, 각 리뷰 키워드가 긍정의 단어인지, 부정의 단어인지를 단번에 알 수 없으며, 모든 키워드가 시각화되어 있어서 콘텐츠에 대한 사용자의 리뷰를 한눈에 파악하기 어렵다는 한계가 존재한다.

본 연구에서는 TF-IDF를 활용해 각 콘텐츠 리뷰 데이터에서 핵심 키워드를 추출하고, 해당 키워드를 기준으로 평점 정보를 재구성했다. 이를 활용하여 콘텐츠별 핵심 키워드 및 각 키워드를 기준으로 재구성한 평점 정보를 시각화하여, 사용자가 각 콘텐츠의 사용자 리뷰를 더 쉽고 자세하게 이해할 수 있는 서비스를 구현하고, 사용자 스터디를 통해, 본 연구의 효과를 검증했다.

2. 관련 연구

콘텐츠의 성장과 더불어 콘텐츠 리뷰에 관한 연구도 활발히 수행되고 있다. 연구[3]에서는 출현 빈도가 높은 단어를 중심으로 영화 관람객의 반응을 시각화했고, 연구[4]에서는 다량의 리뷰 중에서 유의미한 리뷰만을 분석하기 위해 형태소 분석을 진행 후, 추출한 명사의 수를 근거로 유의미한 리뷰를 선별했다. 위 연구와 같이 단순 빈도수를 활용하면 연구 결과에 영향을 주고 있다. 이를 보완하기 위해 특정 문서에서만 자주 등장하는 단어의 중요도를 높다고 판단하는 TF-IDF를 활용한 연구들 또한 많이 진행되고 있다. 연구[5]에서는 뉴스 기사를 요약하여 보여주는 방법으로 TF-IDF를 활용하여 키워드를 추출했다. 연구[6]에서는 소설의 주제어를 추출하기 위해 TF-IDF를 활용했다.

단순 빈도수, TF-IDF 등의 다양한 기법을 활용하여 콘텐츠 리뷰를 분석하고 시각화한 연구도 활발히 수행되고 있다. 네트워크 이미지를 통해서 키워드를 시각화하여 제공하거나[3], 워드크라우드를 통해 키워드를 시각화하여 제공한다 [7, 8]. 네트워크 이미지를 통해 시각화한 경우 네트워크 이미지를 처음 접하는 사용자의 경우 이를 한 번에 이해하기 쉽지 않고, 워드크라우드를 통한 시각화는 여러 개의 키워드만을 제공한다.

본 연구에서는 특정 콘텐츠에서만 의미있는 키워드를 추출하여 사용자가 더 빠르게 리뷰를 이해하고 평점을 뒷받침할 근거로 사용할 키워드가 필요했다. 따라서, 모든 작품에서 자주 등장하는 단어는 중요도가 낮다고 판단하며, 특정 작품에서만 자주 등장하는 단어는 중요도를 높게 판단하는 TF-IDF를 사용했다. 또한 사용자가 이미지를 한 번에 이해하고, 키워드와 키워드별 점수를 제공하기 위해 방사형 그래프를 통해 시각화했다.

3. 연구 방법

3.1 데이터 수집

사용자의 리뷰들 속에서 키워드를 추출하여 키워드와 함께 점수를 제공하기 위해 리뷰별 점수들이 존재하는 형태의 정보가 필요했다. 리뷰 데이터 수집을 위해, 웹 브라우저를 제어할 수 있게 해주는 라이브러리인 'Selenium'을 사용하여 '키노라이츠'[9]에서 17개 작품의 리뷰 데이터 및 등장인물 정보를 크롤링했다.

콘텐츠 평점 분석 웹서비스를 구현하기 위해 필요한 정보 중 작품명, 개봉 일자, 총 평점, 총 2061개(평균 121개)의 점수가 존재하는 리뷰, 리뷰별 점수 등을 추출하여 DB에 저장했다.

3.2 형태소 및 키워드 분석

키워드를 추출하기 위해 문장을 단어로 나누는 과정이 필요했다. 하지만 모든 단어보다는 의미를 가지는 단어들이 필요했기 때문에 KoNLPy 형태소 분석 라이브러리의 Open Korean Text 형태소 분석기를 사용했다. 명사, 동사, 형용사, 부사 총 4가지 품사의 형태소에 해당하고 두 글자 이상의 단어를 작품별로 텍스트 파일에 저장했다. 이 과정에서 작품명과 등장인물의 정보는 불용어로 정의했다. TF-IDF는 TF와 IDF를 곱한 값으로 TF는 특정 문서에서의 특정 단어의 등장 횟수, IDF는 특정 단어가 등장한 문서의 수에 반비례하는 수이다. 작품별로 단어를 저장한 텍스트 파일을 하나의 문서로 정의하고 TF-IDF 값을 구했다. 작품 당 TF-IDF 값이 높은 5개의 키워드를 추출하고 해당 키워드가 포함된 각 리뷰의 점수 평균을 계산하여 표[1]과 같이 키워드별 리뷰 평균 평점 점수를 산출했다.

표[1]. 영화 택배기사 키워드 추출

키워드	리뷰 평균 평점
김우빈	3.4
매드맥스	1.8
세계관	2.6
영웅	1.6
지구	2.5

3.3 구현 내용

데이터 수집 과정과 키워드 분석 단계에서 DB에 저장한 정보들을 이용해 그림[1]과 같이 웹서비스를 구현했다. DB에 있는 데이터들을 사용자에게 제공하기 위해서 Django를 사용했다. 콘텐츠 평점 분석 웹서비스의 주된 첫 번째 기능인 '평점 세분화'는 JavaScript로 방사형 그래프를 시각화하여 키워드별 점수를 표시해 주었다. 두 번째 기능인 '키워드별 리뷰'는 해당 키워드 버튼을 누르면 키워드가 포함된 리뷰들을 평점 3점을 기준으로 3점 이상은 '좋아요' 섹션에, 3점 미만은 '아쉬워요' 섹션에 표시해 주었다.

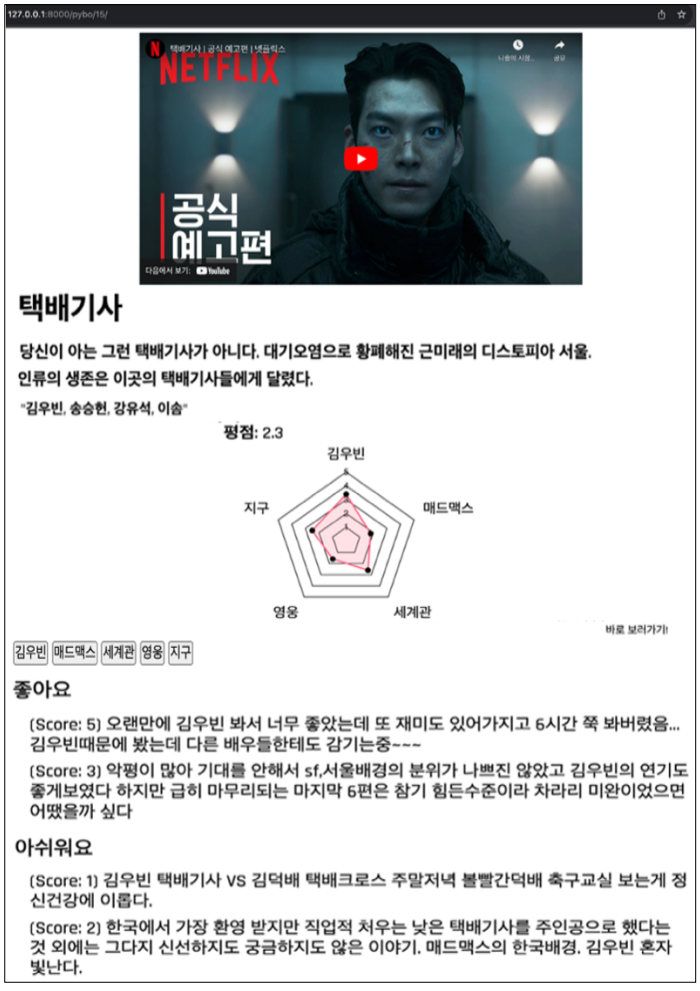

그림 1. 영화 [택배기사] 상세페이지 및 키워드 및 김우빈 키워드에 대한 리뷰 필터링 결과

4. 사용자 스터디

대학생 및 직장인 4명(남자2, 여자2)을 대상으로 사용자 스터디를 진행했다. 4명의 사용자에게 리뷰 사이트 사용 경험에 대해 질문한 후, 본 연구에서 구현한 웹서비스에서 원하는 콘텐츠를 자유롭게 검색해 보도록 했다. 체험 후에는 방사형 그래프를 통해 제공된 핵심 키워드가 해당 작품을 설명하거나 이해시키는데 어떤 도움을 주었는지, 키워드 버튼을 통해 해당 키워드가 포함된 리뷰를 볼 때 어떤 도움이 되었는지, 기존에 있던 리뷰 사이트와 비교했을 때 차별화된 장점에 대해서 질문했다.

방사형 그래프와 관련된 답변으로는 4명의 참여자 모두 '키워드를 통해 영화의 내용과 장단점을 간결하게 알려준다.'라는 것에 동의했다. 대표적으로 “많은 시간을 소비하지 않고도 영화 관람의 여부를 결정해 줄 수 있는 지표가 되어줄 수 있을 것 같아요.”, “타 사이트에서 평점만 있고 리뷰가 없을 때 사람들이 어떤 기준으로 점수를 부여했는지 알 수 없지만 방사형 그래프는 키워드별로 점수가 있어서 평점이 아닌 키워드로 작품을 설명할 수 있게 알려주는 것 같아요.”가 있었다. 키워드별 리뷰와 관련해서는 4명의 참여자 모두 '영화에 대한 다양한 감정과 이유를 이해하고, 다른 사람의 의견에 공감하며 원하는 유형의 리뷰를 선택해 볼 수 있다.'라는 것에 동의했다. 대표적으로는 “키워드를 누르면 해당 리뷰를 좋았던 리뷰와 별로였던 리뷰로 나눠서 볼 수 있는 점이 사람들의 다양한 반응을 살펴볼 수 있게 해서 좋았어요.”, “모든 리뷰가 아닌 그중에서 좀 더 제가 궁금한 부분을 사람들이 어떻게 생각하는지만을 볼 수 있는 부분이 영화를 선택할 때 도움이 될 것 같아요.”라고 답변했다. 이는 기존에 제공하고자 했던 서비스의 의도를 만족하며 사용했다는 긍정적인 피드백을 얻을 수 있었다.

5. 결 론

콘텐츠의 성장과 리뷰의 중요성이 대두되는 가운데 사용자가 리뷰를 통해 얻을 수 있는 데이터는 많지 않았다. 본 연구에서는 사용자에게 다양한 데이터를 시각화하여 제공함으로써 영화에 대한 이해도를 높이는 것을 목표했다. 이를 위해 TF-IDF를 활용하여 키워드를 추출하고 키워드별 점수와 함께 방사형 그래프로 시각화하여 제공했다. 또한 키워드별 리뷰를 제공함으로써 영화에 대한 이해도를 높여주었다. 향후 연구에서는 형태소 분석에서의 정확도를 높이고 더 많은 콘텐츠의 리뷰를 활용할 예정이다. 이를 통해 의미 있는 키워드를 추출하고 다양한 콘텐츠를 제공하여 사용자가 더 좋은 경험을 할 수 있도록 발전할 수 있다.

참고문헌

[1] Netflix to Announce Second Quarter 2023 Financial Results. Netflix Investors.
[2] 서민석 외 3인. ABSA 기반 영화 리뷰 감상 포인트 분석 시스템. 2022년
[3] 김슬기 외 1인 의미연결망 분석을 활용한 영화 리뷰 시각화. 2019년
[4] 정지훈 외 2인. 텍스트마이닝 기법과 ARIMA 모형을 활용한 배달의 민족 앱 리뷰 분석. 2021년
[5] 이성직 외 1인. tf-idf의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. 2009년
[6] 유은순 외 2인. tf-idf와 소설 텍스트의 구조를 이용한 주제어 추출 연구. 2015년
[7] 최혜선 외 1인. 밀키트 제품 리뷰 데이터를 이용한 텍스트 분석 사례 연구. 2022년
[8] 김문기. 모바일 서비스 산업의 이용자 만족/불만 요인 탐색에 관한 연구: 모바일 앱 리뷰 분석을 중심으로. 2023년
[9] https://www.kinolights.com