멀티 모달 딥러닝을 활용한 웹소설 추천 시스템

김미려, 김현희
동덕여자대학교 정보통계학과

20181043@dongduk.ac.kr, heekim@dongduk.ac.kr

Multi-Modal Recommendation System for Web Novels

Mi Ryeo Kim, Hyon Hee Kim
Dept. of Statistics and Information Science, Dongduk Women's University

요 약

웹소설 시장의 성장에 따라 웹소설 추천 시스템의 중요성이 높아지고 있다. 본 연구에서는 작품의 특성 및 선호도를 나타낼 수 있는 다양한 데이터를 활용하여 추천시스템을 구현하고 그 성능을 평가하여 표지 이미지와 작품 특성을 모두 고려한 멀티 모달 추천 시스템이 가장 효율적임을 보여주었다. 연구 결과, 단일 변수 추천에서는 작품 소개글과 표지 이미지 기반 추천이 가장 좋은 성능을 보였고, 멀티 모달 추천 시스템에서는 작품 소개글, 이미지, 키워드 순으로 성능에 좋은 영향을 끼치는 것으로 나타났다. 이번 연구 결과는 한국콘텐츠진흥원에서 조사한 웹소설 이용자 실태조사와는 조금 다른 결과를 보여주었다. 설문조사에서는 인기도를 웹소설 선택 시 가장 중요한 영향으로 봤으나, 본 연구에서는 작품 소개글이 가장 중요한 영향을 미친다는 결과가 나타났다. 이러한 연구 결과는 웹소설 추천 시스템의 개발과 운영에 있어서 중요한 참고 자료가 될 것으로 예상된다.

1. 서론

현재 웹소설 시장은 인터넷과 모바일 기술의 발달로 급격하게 성장하고 있다. 한국출판산업진흥원에서 발간한 보고서[1]에 따르면 2020 년 기준 웹소설 산업 규모는 약 7,415 억원이며 높은 성장 전망을 보인다. 또한 웹소설 카테고리는 아시아, 미국을 중심으로 해외로도 사업이 크게 확장되고 있으며, 드라마, 영화 등 웹소설 IP 를 바탕으로 한 2 차 창작물도 큰 수익을 창출하며 '황금알을 낳는 거위'의 면모를 보이고 있다.[2] 하루에도 수십개의 웹소설이 런칭되는 시장 환경에서 독자들은 취향의 맞는 소설을 찾기 위해 많은 노력을 기울어야 한다. 이에 웹소설 추천 시스템의 중요성이 높아지고 있기 때문에, 본 연구는 인기도, 표지 이미지, 작품 소개글 등 다양한 멀티 데이터를 이용하여 콘텐츠 기반 웹소설 추천 시스템을 개발하고자 한다.

본 연구에서는 추천 시스템을 구축하는데 필요한 변수들을 선정하기 위해서 한국콘텐츠진흥원에서 2020 년 발간한 웹소설 사용실태조사[3]를 참고하여 변수를 설정했다. 가장 많이 이용하는 플랫폼 설문과 웹소설 선택 기준 설문을 참고하여 카카오페이지를 주 플랫폼으로 선정하였으며, 높은 비율로 선택된 '인기 순', '소재, 줄거리', '가격'을 변수로 선정했다. 세밀한 추천을 위해 장르는 '로맨스 판타지' 장르로 한정했다. 또한 높은 응답을 보이는 것은 아니나, 상품 페이지에서 웹소설의 표지가 함께 보여지기 때문에 표지 역시 변수로 설정했다.

다음으로, 설정한 추천 변수의 임베딩 값을 이용하여 코사인 유사도 행렬을 만든 뒤, 단일 추천 모델을 구축하였다. 또한 멀티 모달 추천 시스템에서 중요한 변수의 값을 찾기 위해, 1001 가지의 경우의 수로 각 변수의 임베딩 값에 가중치를 두어 멀티 모달 추천 시스템을 구축했다. 각 변수에 가중치를 두기 위해 다양한 조합으로 실험을 실시했으며, 실험 결과 가장 좋은 성능을 보였던 가중치 조합을 선택했다. 단일 추천 모델에서 성능이 좋았던 작품 소개글과 표지 이미지의 가중치가 높았을 때, 멀티 모달 추천 시스템을 구축하였을 경우, 가장 좋은 성능을 보였음을 확인했다. 이러한 결과는 웹소설 추천 시스템의 개발과 운영에 있어서 중요한 참고 자료가 될 것으로 기대된다.

본 논문은 다음과 같이 구성된다. 제 2 장에서는 변수 선정과 수집 및 전처리에 대해 다룬다. 제 3 장에서는 2 장에서 설정한 변수를 활용한 단일 추천 모델

을 제시한다. 제 4 장에서 단일 모델의 임베딩 값에 가중치를 주어 구현한 멀티 모달 추천 시스템을 제시하며 마지막으로 결론 및 향후 연구를 제시한다.

2. 데이터 수집 및 전처리

2.1 변수 선정

먼저 가장 중요한 요인으로 판단되는 '인기순' 변수는 단순히 랭킹의 순위로 보지 않고 인기도로 확장하여 변수로 활용하였다. 카카오 페이지 '월간 랭킹 300 위 포함 여부', '별점', '조회수 및 댓글수', '원작 IP 를 활용한 2 차 저작물 여부'(본 연구에서는 '웹툰화'라고 지칭한다.), '단행본 발간 여부'도 인기도 변수로 선정했다.

'소재 줄거리'는 '작품 소개글'과 '키워드', '제목'을 통해 수집했다. 웹소설의 경우 소설 내용이 유추가 가능하도록 제목을 짓는 성향이 있어 '제목' 역시 변수에 포함시켰다.

이외에도 '완결된 정보를 한 번에 몰아보는 것을 선호한다'는 질문에 대해 긍정 응답률이 74.4%, '기다렸다가 무료로 이용하는 것을 선호한다'에 대해 그렇다는 긍정적인 응답의 비중이 57.4%로 나타났기 때문에 '완결 여부', '기다리면 무료 여부'를 추가 변수로 설정했다.

2.2 데이터 수집 및 전처리

'키워드'를 제외한 위의 변수들을 Python 의 Selenium 을 이용하여 2023 년 1 월 28 일 크롤링하였다. 주 플랫폼으로 설정한 '카카오 페이지에서'는 키워드 대신 유저 반응을 제공하는데, 유저 반응은 '울뻔한', '고귀한'과 같이 추상적이므로 타 플랫폼인 '리디'에서 키워드 정보를 크롤링했다. '카카오 페이지'에서 제공하는, 19 세 이상 이용 가능 작품이 아닌 로맨스판타지 웹소설 5457 개의 작품 중 '리디'에서 키워드를 제공하며, 키워드 전처리 결과 키워드가 한 개만 남은 70 작품을 제외한 **총 2966 개의 작품**을 분석 데이터로 확보했다.

전처리 작업은 다음과 같이 진행하였다. 먼저, 다중공선성 값을 구해 Vif 값이 5 이상인 '댓글수'를 변수에서 제외했으며, '조회수'와 '별점'을 alpha 값이 0.7 인 휴리스틱 스코어링 하고 변수 이름을 'score'로 정했다.

다음으로 '키워드'의 경우, 작품의 소재에 대한 내용이 아닌, 별점과 가격에 대한 키워드를 제외, 총 143 개의 범주의 키워드를 사용했다. 다만 출간 년도에 관한 키워드는 겹치는 변수가 없고 웹소설은 유행이 빠르기 때문에 포함했다.

3. 단일 추천 모델

각 추천 시스템마다 임베딩 값을 구한 뒤, 구한 임베딩 값을 이용해 코사인 유사도 행렬을 만들었으며, 성능평가는 NDCG 로 진행했다. Relevant item 은 평균 score 값보다 큰 값으로 설정했고, 각 추천의 평균 NDCG 값을 구해 추천 시스템의 성능 평가 지표로 삼았다.

가장 기본적인 추천 모델로는 작품 기본 정보 및 인기도 정보를 사용하였으며 이를 위해 '전체화수'에 z-score 정규화를 적용하고, '기다리면무료여부', '연재정보', '단행본여부'와 같은 범주형 변수에는 원핫인코딩을 적용해 임베딩 값을 구했다.

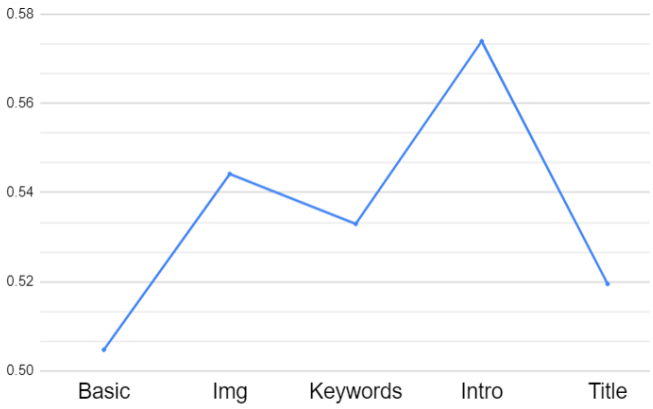
표지 이미지 활용을 위해 VGG16[4]을 이용한 전이학습을 통해 표지 이미지를 임베딩 값을 구하였으며 제목을 활용하기 위해서 키워드를 활용하였다. 이를 위해 제목의 문장 유사도가 아닌, 사용한 단어의 유사도를 구하자 Word2Vec[5]을 사용하여 임베딩 값을 구했다. 또한 작품 소개글을 활용하였는데 한 문장으로 끝나는 제목과 달리, 작품 소개글은 여러 문장으로 이루어져 있기 때문에, 문장 유사도를 구하기 위해 Doc2Vec[6]을 사용하여 임베딩 값을 구하였다.

키워드를 활용하기 위해서 sklearn 에서 제공하는 Countvectorizer[7]를 이용해 키워드 임베딩 값을 사용하였다. 143 개의 범주로 정리된 키워드를 데이터로 사용했고, 소재의 의미를 고려하여 제목 기반 추천을 진행했기 때문에 키워드의 경우 단순히 출현 빈도를 기반으로 임베딩을 진행했다.

최적의 추천 개수를 찾기 위해, 추천 개수를 [2, 3, 5, 7, 10, 15, 20, 25, 30, 35, 40] 바꿔가며 위의 기본정보 및 인기도 기반(Basic), 표지 이미지 기반(Img), 제목(Title), 작품소개글(Intro), 키워드(Keywords)기반 추천 시스템의 NDCG 값을 구해 비교해보았다. 그 결과 추천 개수를 15 개로 정했다.

단일 추천 모델의 결과는 그림 1 과 같다. 그림 1 에서 볼 수 있는 바와 같이 작품 소개글을 활용한 추천 모델이 가장 높은 성능을 보였으며, 순차적으로 표지 이미지, 키워드, 타이틀 변수로 나타났다.

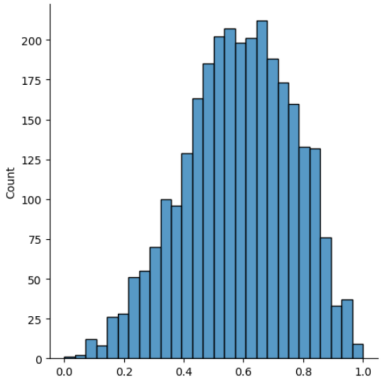
(그림 1) 단일 추천 모델의 NDCG 값 비교



4. 멀티 모달 추천 시스템

기본 정보 및 인기도, 표지 이미지, 제목, 작품소개글, 키워드를 고려한 멀티 모달 추천 시스템을 구축하였다. 코드를 통해 랜덤하게 생성한 1001 가지 경우의 수로 각각의 임베딩 값에 가중치를 정의하였으며, 모든 임베딩 값을 합쳐 추천을 실시하였다. 그 결과, 작품 소개글 임베딩 값에 0.7 가중치를 표지 이미지 임베딩 값에 0.2 가중치를, 키워드 임베딩 값에 0.1 가중치를 주었을 때 멀티 모달 추천 시스템의 성능이 가장 좋았다. 단일 추천 모델에서 가장 높은 성능을 보였던 작품 소개글 기반 추천 모델과 비교해보면 NDCG 값이 약 0.1 상승했음을 알 수 있다.

(그림 3) 멀티 모달 추천 시스템의 NDCG 평균 값 분포



위 그래프는 본 연구에서 구현한 멀티 모달 추천 시스템의 NDCG 분포도로, 각 소설에 대한 추천 결과의 NDCG 값의 분포를 보여준다. NDCG 값의 분포는 오른쪽으로 치우쳐진 정규분포의 형태로 추천 시스템의 일반화 능력이 높고, 안정적인 성능을 가진다는 점을 알 수 있다.

(표 1) 멀티 모달 추천 시스템 결과

Similarity: 0.583

<기준작>

<추천작>

5. 결론 및 향후 연구

본 연구에서는 웹소설을 고르는 기준이 되는 변수를 설정하고 각 변수를 사용한 콘텐츠 기반 단일 추천 시스템을 구현했다. 그 뒤, 각 변수의 임베딩 값을 합쳐 텍스트 및 이미지 데이터를 고려한 콘텐츠 기반 멀티 모달 추천을 진행했다. 그 결과, 단일 모델에서는 작품 소개글과 표지 이미지 기반의 추천 시스템의 성능이 가장 좋았으며, 인기도 및 작품의 기본 정보를 기반으로 한 추천 시스템의 성능이 타 추천 시스템보다 떨어짐을 알 수 있었다. 또한 멀티 모달 추천 시스템을 구현할 때 작품 소개글, 이미지, 키워드 순으로 성능에 좋은 영향을 미치는 것을 알 수 있었다.

본 연구 결과는 한국콘텐츠진흥원에서 조사한 웹소설 이용자 실태조사와는 조금 다른 결과로, 사용자들의 설문에서는 인기도가 웹소설 선택 시 가장 중요한 영향을 끼친다고 나왔으나 연구 결과 줄거리에 해당하는 작품 소개가 가장 중요한 영향을 끼침을 알 수 있었다. 또한 표지, 이미지의 경우 설문에서는 크게 두드러지지 않았으나, 연구 결과 높은 영향을 미치는 것을 알 수 있었다. 이는 표지의 경우 작품의 분위기가 간접적으로 드러나며 주인공의 외양 묘사가 들어가는 경우가 많아 높은 영향을 주는 것으로 생각된다.

웹소설은 플랫폼 특성상 한 편씩 결제되는 구조로 구매 내역을 통해 독자의 호불호를 확실하게 알 수 있다. 본 연구는 콘텐츠 기반 추천 시스템으로 모든 독자에게 동일한 웹소설을 추천해준다. 따라서 이후 연구에서는 사용자의 구매 이력 및 평가 정보를 활용하여 개인 맞춤형 추천 시스템을 개발하고자 한다.

이러한 연구를 통해 웹소설 시장에서 개인화된 추천 시스템의 중요성을 확인하고, 이를 통해 독자에게 보다 만족스러운 서비스를 제공할 수 있는 방안을 모색하고자 한다.

참고문헌

[1] 한국출판산업진흥원. (2021). [조사연구 2021-04] 전자출판 산업분석 및 활성화를 위한 조사연구 보고서

[2] Nielsen. (2020, August26). IP 무한 확장의 시대, 웹툰/웹소설 시장 왕좌의 주인은?. Korean Click, 307-2.

[3] 한국콘텐츠진흥원. (2020). 2020 웹소설 이용자 실태조사

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] 위와 동일.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.