고객 맞춤형 웹툰 추천 시스템 제안

권송아(소프트웨어학부) , 신수현(수학과), 이상화(소프트웨어학부), 이하은(소프트웨어학부)

2021 CUAI 중앙대학교 인공지능 학회 하계 컨퍼런스
Proceeding of 2021 Chung-Ang University Artificial Intelligence's Summer Conference

Abstract

- 문화 콘텐츠 중 웹툰에 초점을 맞추고 사용자가 좋아할 만한 웹툰을 추천하는 서비스를 개발.
- 추천 방식은 콘텐츠 기반 추천에 스토리, 그림체 기반 알고리즘, 협업 필터링에 아이템, 잠재 요인 기반 알고리즘로 나뉨.

Introduction

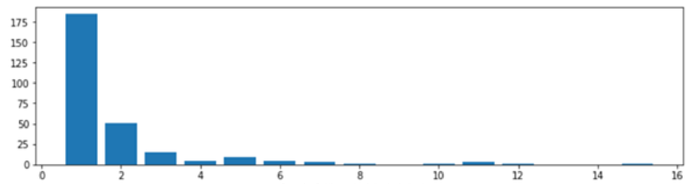
- 기존 웹툰 플랫폼에서 제공하는 추천 시스템은 하나의 플랫폼 안에서만 결과를 제공하며 추천 요소가 다양하지 않다는 단점이 있음. 이러한 부분을 보완하고자 대표적인 웹툰 플랫폼들의 웹툰을 통합하고, 사용자가 필요한 요소들에 따라 추천 알고리즘을 각각 개발하여 결과를 제공할 것을 제안
- 추천 시스템은 콘텐츠 기반과 협업 필터링을 이용한 추천으로 나뉘는데, 콘텐츠 기반 추천은 웹툰이 가진 속성에 기반하는 것으로 실험에 사용된 속성은 스토리와 그림체가 있음. 협업 필터링을 이용한 추천은 각 사용자의 선호도를 반영한 것으로 아이템과 잠재 요인에 기반하여 설계.
- 각 추천 알고리즘의 정확도를 높이기 위해 여러 실험을 진행하였고, 이에 따른 결과를 정리하여 최종 추천 시스템을 제안

Aim

- 사용자에게 여러 플랫폼의 웹툰을 통합 추천하여 넓은 선택의 폭을 제공.
- 콘텐츠 기반 추천 시스템과 사용자 행동 양식 기반 추천 시스템 알고리즘을 사용하여 다양하지 못했던 기존 플랫폼의 추천 시스템을 보완.

Materials

- 네이버 웹툰, 다음 웹툰, 카카오페이지, 레진코믹스에서 약 7400 여개의 웹툰 데이터를 크롤링하여 활용.
- 사용자 104명에게 선호웹툰장르와 선호/비선호 웹툰을 조사. 선호웹툰장르로는 '일상', '개그', '판타지', '드라마' 등을 제시 및 콘텐츠 기반 추천 시스템에서 사용. 응웹툰은 선호 웹툰이 277개, 비선호 웹툰이 51개, 중복 제외 총 303개. 사용자 별 평균 응답 웹툰 수는 6.44개였으며, 웹툰 별 사용자의 응답 수는 1개인 경우가 184개로 전체 응답의 절반 이상을 차지함. 해당 데이터는 사용자 행동 양식 기반 추천 시스템에서 사용



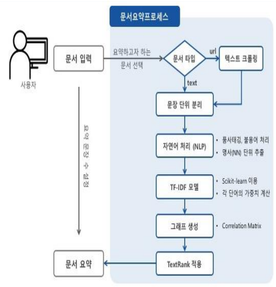
[그림 1] 웹툰 별 사용자의 응답 수

Methods

1. 콘텐츠 기반 추천 - 스토리 기반

$$TR(V_i) = (1-d) + d * \sum_{V_j \in In(V_j)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} TR(V_j)$$

- $TR(V_i)$는 단어 (V_j)의 TextRank값, W_{ji}는 단어 i와 j사이의 가중치. d는 damping factor로 PageRank알고리즘에서 사용자가 해당 페이지를 만족하지 못하고 다른 페이지로 이동할 확률. TextRank 알고리즘에도 그대로 적용.
- 수식을 통해 각각의 단어의 가중치를 계산 후, 높은 순으로 정렬하면 주어진 문장에서 중요 단어를 확인 가능.



[그림 2] TextRank 알고리즘 프로세스

- TF-IDF 모델은 특정 문서 내에서 단어 빈도를 통해 해당 문서 내 단어의 가중치를 계산하여 핵심어를 추출하는 알고리즘
- 웹툰 줄거리에서 Ranking 값이 높은 순으로 정렬하여 3개의 단어 추출 및 키워드 생성.
- 이 때, 미사여구, 대명사, 연결어구 등을 stopword라는 카테고리를 이용하여 제거한 후 추출된 단어를 대상으로 TF-IDF 계산 및 가중치 그래프를 만들고, 이에 TextRank 알고리즘을 적용하는 방식

2. 콘텐츠 기반 추천 - 썸네일 기반

[그림 3] 네이버 웹툰 "대학일기"의 썸네일 이미지에 대한 모델 별 유사도 측정 결과
(왼쪽부터 차례대로 MobileNetV3-Small, ResNet18, ResNet50)

- 썸네일 dataset을 대상으로 RGB 평균과 표준편차값을 다시 계산하여 정규화 및 각 이미지의 크기를 125*125로 일정하게 설정 후 여백은 흰색으로 패딩
- 코사인 유사도를 계산한 값을 바탕으로 결과 도출. MobileNetV3-Small, ResNet18, ResNet50 세 모델을 정확도와 소요 시간을 기준으로 비교했을 때, ResNet18의 성능이 가장 우수하다 판단.

3. 사용자 행동 양식 기반 추천 - 아이템 기반 협업 필터링

- 각 아이템에 대한 사용자들의 행동 양식을 기반으로 아이템 간의 유사도를 계산하는 방식.
- 응답 받은 선호웹툰 데이터로 선호도 행렬을 만들었고, 이를 사용자 단위로 비교해 코사인 유사도 계산. 코사인 유사도 이외에 유클리디안, 자카드 유사도도 고려했으나 코사인 유사도가 사용한 데이터 세트에 가장 적합하다고 판단.

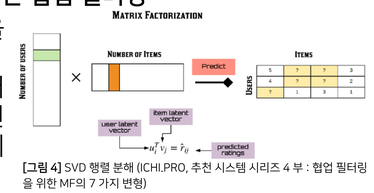
$$\hat{P} = (P_{u,k} * S_{k,i}) / \sum_i (|S_{k,i}|)$$

- $\hat{P}$: 사용자 u, 아이템 i에 대한 예측 선호도
- $P_{u,k}$: 사용자 u가 아이템 k에 대해 평가한 실제 선호도
- $S_{k,i}$: 아이템 k와 아이템 i 사이의 유사도
- $\sum_i (S_{k,i})$: 정규화를 위한 term

- 최종 예측 유사도는 사용자-웹툰 선호도 행렬과 코사인 유사도를 곱한 후, 각 아이템에 대한 코사인 유사도 벡터의 합을 나누어 계산.

4. 사용자 행동 양식 기반 추천 - 잠재요인 기반 협업 필터링

- 사용자-아이템 평점 행렬 속 잠재 요인을 추출하여 추천을 예측하는 기법.
- 잠재 요인을 기반으로 다차원 행렬 데이터를 두 개의 저차원 행렬로 분해 및 분해된 두 행렬 내적을 통해 새로운 예측 행렬 데이터 생성.



[그림 4] SVD 행렬 분해 (ICHI.PRO, 추천 시스템 시리즈 4 부 : 협업 필터링을 위한 MF의 7 가지 변형)

- 사용자-K차원 잠재요인 행렬(P)과 K차원 잠재요인-아이템(Q.T)의 내적은 차원으로 구성된 사용자-아이템 행렬(R) (이 때, 아이템은 웹툰이고, 사용자는 웹툰을 시청하는 이용자). R 행렬의 1행 사용자와 2열 웹툰에 해당하는 값은 미정 데이터이지만, $r_{(1,2)} = p_1 \times q_2^t$으로 유추 가능.
- R 행렬과 예측 행렬(P · Q.T)의 RMSE 계산 시 오류가 가장 적게 나온 파라미터를 바탕으로 모델 수행. 각 파라미터 값은 잠재 요인의 차원수(K)는 8, SGD의 반복 횟수(steps)는 200, 학습률(learning_rate)은 0.006, 규제계수(r_lambda)는 0.01로 설정.

Results

1. 콘텐츠 기반 추천 - 스토리 기반

```
title :  다시피는 꽃     keywords : ['이야기', '위안부', '할머니']
title :  트리니티 원더   keywords : ['고수', '마법사', '무협']
title :  피그말리온 컴플렉스     keywords : ['고아', '변신', '소녀']
title :  오늘의 초능력   keywords : ['소녀', '불허', '예측']
title :  골든왈츠        keywords : ['기적', '사랑', '아이스댄싱']
```

[그림 5] 웹툰 별 키워드 추출

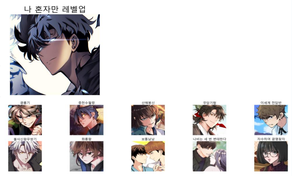
- 각 웹툰의 키워드를 추출 후, 전체 웹툰 코사인 유사도를 비교
- 유사도가 높은 순으로 상위 10개 추출

각자의
[['2015', 0.795064394779917, '네이버 웹툰'], ['너에게 돌아가는 길', 0.63591590266479, '레진코믹스'], ['바로크 기사단', 0.590063737877753
26, '카카오 페이지'], ['옆집남자', 0.5821390895208248, '다음 웹툰'], ['소년들은', 0.5737335336450612, '네이버 웹툰'], ['빠져드는 한숨에
달콤한 키스', 0.5400281717342458, '레진코믹스'], ['ZZIM! (찜!)', 0.5400281717342458, '레진코믹스'], ['단 한가지의 룰', 0.53305985487621
27, '레진코믹스'], ['이것은 사랑 이야기', 0.5308244583384923, '레진코믹스'], ['685일', 0.5249179759499697, '레진코믹스']]

[그림 6] 유사웹툰 상위 목록

2. 콘텐츠 기반 추천 - 썸네일 기반

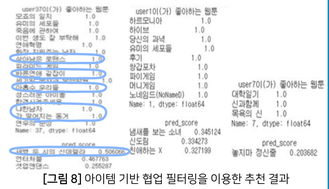
- ResNet18을 사용하여 사용자에게 보여줄 결과를 출력. 유사도가 높은 순서대로 0.8 이상인 웹툰 20개 결정 및 사용자가 선호하는 장르의 웹툰부터 유사도가 높은 순서대로 제시.
- 이때, 각 웹툰의 제목과 썸네일 이미지를 출력



[그림 7] 그림체 기반 추천 결과

3. 사용자 행동 양식 기반 추천 - 아이템 기반 협업 필터링

- 로맨스물을 좋아하는 사용자들([그림8]의 user37)에게는 로맨스물이 추천됨.
- 로맨스물을 좋아하지 않는 사용자([그림8]의 user1, user7)에게는 로맨스물이 아닌 웹툰이 추천됨.



[그림 8] 아이템 기반 협업 필터링을 이용한 추천 결과

4. 사용자 행동 양식 기반 추천 - 잠재요인 기반 협업 필터링

	pred_score	user_id
바른연애 길잡이	1.165272	2
파라미드 게임	1.071151	2
재혼 황후	1.060038	2
소녀의 세계	1.044844	3
여주살격!	0.886652	3
전지적 독자 시점	0.836209	3
세기말 풋사과 보습학원	0.997757	8
마음의소리	0.875702	8

[표 1] 잠재 요인 협업 필터링 예측 결과

- 즉각적인 피드백이 없어 추천 시스템 결과에 대한 완벽한 평가와 대처가 이루어지지 않았지만, 사용자들의 선호하는 웹툰 장르와 비슷하게 웹툰이 추천됨.
- 'user_id'가 3이 선호하는 장르인 판타지를 기반으로 한 " 전지적 독자 시점" 웹툰이 추천됨.

Conclusion

- 사용자의 스토리 및 그림체 취향과 웹툰 선호도를 통해 웹툰을 추천하는 시스템
- 여러 추천 시스템을 통해 사용자의 상황에 맞는 개인화된 추천을 제시 및 네이버, 카카오페이지 등 주요 웹툰 플랫폼을 통합한 데이터를 사용해 한 플랫폼에 국한되어 있던 기존 서비스의 범위를 넓혀 사용자에게 폭 넓은 선택지를 제공.
- 스토리 기반 추천은 줄거리에서 추출한 키워드 사이에서 유의어를 정확히 식별하지 못하는 문제가, 그림체 기반 추천에서는 모든 썸네일이 그림체를 대변하는 것은 아니라는 문제가 존재
- 협업 필터링 기반 추천은 수집한 데이터가 부족해 추천에 대한 정확도가 떨어지며, 사용자 피드백의 부재로 정확도가 떨어지는 한계점 존재.
- 이러한 한계점을 보완하고, 대규모의 데이터 처리에 적합하도록 알고리즘을 개선시키면 더 나은 추천이 이루어질 것이라 예상함.

Reference

- P. Wongchaisuwat, "Automatic Keyword Extraction Using TextRank," 2019 IEEE 6th International Conference on Industrial Engineering and Applications (ICIEA), 2019, pp. 377-381
- 이지훈, 엄태현, 이혁준. "TF-IDF와 CNN을 사용한 법률문서 분류 시스템에 관한 연구." 한국통신학회 학술대회논문집 . (2019): 982-983.
- 권철민, 파이썬 머신러닝 완벽 가이드, 2020, pp. 567-579, 591-606
- 김형도. "잠재 요인 모델의 원리를 이용한 협업 태그 기반 추천 방법." 한국전자거래학회지 14.4 (2009): 47-57.
- 이미지 탐지기 쉽게 구현하기, 당근마켓 팀블로그