

TF-IDF 및 DeepFM 기반 고객 맞춤형 OTT 콘텐츠 추천 서비스

전택희*, 고은비*, 신수민*, 탁효준*, 박효원*, 박상훈*, 손종수*
*CJ올리브네트웍스

Customized OTT Contents Recommendation Service Using TF-IDF and DeepFM

Jeon, Taek Hee°, Go, Eun bi°, Shin, Su min°, Tak, Hyo jun°, Park, Hyo won°,
Park, Sang Hoon°, Sohn, Jong Soo°
CJ Olivenetworks

E-mail : th.jeon@cj.net, eunbi.go@cj.net, sm.shin3@cj.net, hyojun.tak@cj.net,
hyowon.park@cj.net, sanghoon.park6@cj.net, jongsoo.sohn@cj.net

요 약

최근, OTT 서비스 시장이 확대되고 있으며, 고객에게 맞춤형 콘텐츠를 추천하기 위한 알고리즘의 고도화가 중요한 경쟁력으로 부각되고 있다. 기존 콘텐츠 추천 알고리즘은 전통적인 협업 필터링 및 콘텐츠 기반 추천을 사용하고 있으나 고객의 기호와 취향을 반영한 추천 결과를 제공하기 어렵다. 본 논문에서는 고객의 기호와 취향을 추천 결과에 반영하기 위하여 TF-IDF(Term Frequency-Inverse Document Frequency) 및 Deep Factorization Machine(DeepFM) 알고리즘을 활용한 추천 시스템을 제안한다. 제안된 기법을 적용하여 OTT 서비스의 콘텐츠 추천이 82.3% 정확도를 보이는 것을 확인할 수 있었다.

1. 서론

OTT 서비스는 인터넷을 통해 방송 및 영화, 교육 등 다양한 미디어 콘텐츠를 제공하는 서비스를 의미한다. 최근, OTT 서비스는 단순히 영상 콘텐츠를 재생해주는 기능을 넘어 실시간 콘텐츠 제공 및 개인 맞춤형 콘텐츠 추천 시스템을 기반으로 고객 편의성을 극대화한 서비스를 제공하고 있다.

추천 서비스의 발전과 함께 추천의 정확도는 플

랫폼의 매출 지표를 견인하는 요인으로 작용하고 있다. OTT 전문 기업 넷플릭스(NETFLIX)는 수만 개의 태그 및 개인 취향을 고려하여 그룹화함으로써 개인화 맞춤 추천을 제공한다. 아마존 프라임 비디오(Amazon Prime Video)는 고객의 프로파일, 행동 패턴 파악을 통해 초 개인화된 추천 서비스를 제공하며, 유튜브(YouTube)는 다양한 알고리즘을 기반으로 영상물을 수행하여 추천 서비스를

제공하고 있다. 이처럼 주요 OTT 업체들은 사용자 데이터를 기반으로 추천 서비스 고도화를 통해 서비스 경쟁력을 강화하고 있다[1-3]. 그러나, 국내 OTT 플랫폼은 콘텐츠를 기반으로 모든 사용자에게 획일화된 추천 서비스를 제공하고 있으며 개인화 추천서비스는 미흡하다. 따라서, 본 연구에서는 OTT 서비스 내 개인 맞춤형 콘텐츠를 추천을 제공 하기 위해 TF-IDF 알고리즘을 통해 콘텐츠 설명 분석 및 콘텐츠 간의 유사도를 계산하고, 사용자의 특성을 반영하기 위해 DeepFM 을 적용하고, 특징을 결합하여 OTT 서비스 내 최적화된 사용자 맞춤형 콘텐츠 추천 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 논문에서 사용되는 알고리즘을 소개하고, 3장에서는 프로젝트에 사용된 데이터 및 실험에 대해서 설명한다. 4장에서는 프로젝트에서 얻은 결과를 정리한다. 마지막으로, 5장에서는 논문에서 제안한 방법에 대한 결론 및 향후 방향에 대해서 기술한다.

2. 배경 기술

2.1. TF-IDF

TF-IDF(Term Frequency-Inverse Document Frequency) 모델은 여러 문서 내에서 단어의 중요도를 나타내는 통계적 수치로 정보의 검색 및 텍스트 마이닝 분야에서 널리 사용되고 있다. TF-IDF 에서 단어 빈도 (Term Frequency: TF) 는 특정 단어가 전체 문서에서 나타나는 빈도를 의미한다 [4]. 문서에서 등장하는 빈도가 높은 단어는 중요도가 높다고 예상할 수 있다. 문서 빈도 (Document Frequency: DF)는 하나의 단어가 전체 문서 집합에서 나타나는 빈도를 나타내는 값으로 해당 단어가 등장하는 문서의 개수를 의미한다. DF가 높은 단어는 그 단어가 많은 문서에서 나타났기 때문에 중요한 단어가 아니다. 따라서, DF 값에 역수를 취한 IDF(Inverse Document Frequency)를 사용하여 특정 단어의 중요도를 나타낸다. TF-IDF는 아래의 식 (1)과 같이 계산할

수 있다.

$$TF - IDF_{w,d} = f_{w,d} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \quad (1)$$

위 수식에서 w 는 특정 단어, d 는 특정 문서, D 는 전체 문서를 의미한다. $f_{w,d}$ 는 문서 d 에 출현한 단어 w 의 빈도, $f_{w,D}$ 는 w 가 등장한 문서의 수, $|D|$ 는 전체 문서의 수를 의미한다.

2.2. DeepFM

추천 시스템에서 CTR(Click-Through Rate)를 예측하는 것은 매우 중요하다[5-7]. DeepFM 은 Pre-training, Feature Engineering 이 요구되지 않는 CTR 예측 방법이다[8]. DeepFM 은 임베딩 계층을 통해 희소한 데이터를 압축하여 저차원의 밀집된 데이터로 변환하여 학습에 활용한다. 낮은 차원의 특징 상호작용은 FM에서 학습하고, 높은 차원의 특징 상호작용은 DNN 에서 학습하는 특징이 있다. DeepFM 모델은 FM(Factorization Machine)과 DNN(Deep Neural Networks)의 결합된 모델 구조이며 아래의 식 (2)와 같다.

$$y = \text{sigmoid}(y_{FM} + y_{DNN}) \quad (2)$$

위 수식에서 y_{FM} 는 FM의 출력, y_{DNN} 는 DNN의 출력을 의미한다.

3. TF-IDF 및 DeepFM을 활용한 추천 시스템

본 논문에서 OTT 서비스 콘텐츠 추천을 위해 Candidate Generation and Ranking Model 구조를 기반으로 제안된 방법은 그림 1 과 같다.

총 두 단계로 추천이 이루어지며, 첫 번째 데이터는 OTT 서비스의 프로그램, 장르, 시청 시간 등 16 개의 특징 항목을 포함한 데이터 총 100 만건을 활용하였다. 데이터 전처리를 통해 시청 시간 1 초 이하의 기록들은 학습에 활용하지 않는다. 또한, 하나의 콘텐츠를 두 번 이상 시청

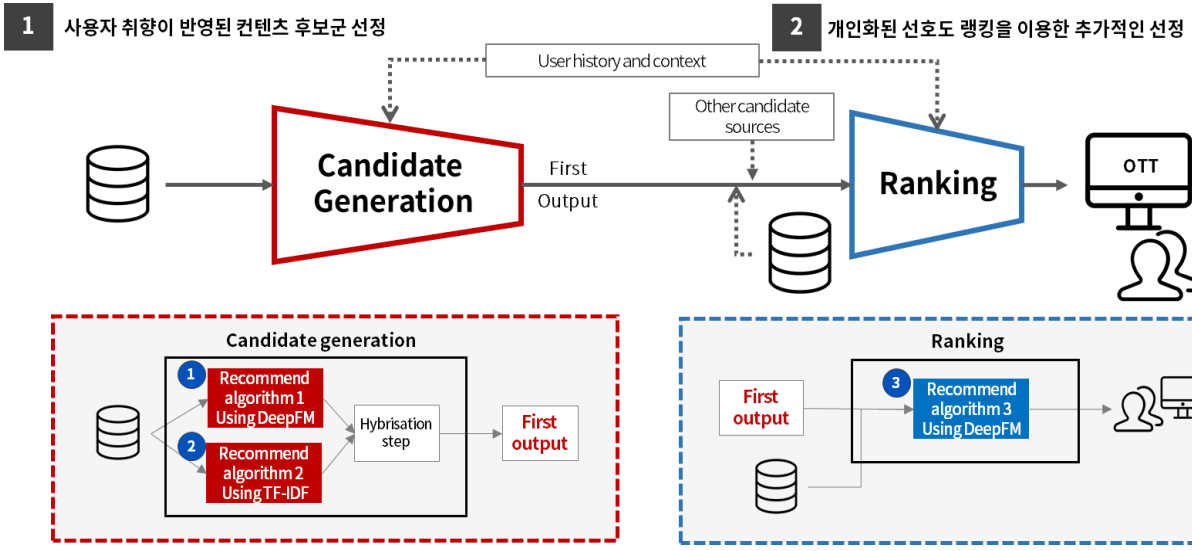


그림 1. Candidate Generation and Ranking Model 기반 OTT 콘텐츠 추천 방법

한 경우 시청 시간 값이 100%를 초과 하기 때문에, 학습 시 발생할 수 있는 편향을 고려 하여 시청 시간을 100% 로 정규화 하였다.

모든 데이터 특징 요소는 표 1 과 같다. OTT 콘텐츠 추천 시스템을 개발하기 위해 python 3.8.8, tensorflow 2.4.1, pytorch 1.8.0 및 KoNLpy 0.5.2 버전을 사용하여 전처리 및 모델 개발, 학습을 수행하였다.

프로그램	프로그램 ID, 러닝타임, 출연진, 제작 년도, 국가(13개), 장르(72개), 시리즈, 중영 여부, 폭력성, 선정성, 외부 판매 여부, 프로그램 설명
개인	연령, 성별, 사용자 Index, 시청 시간

표 1. OTT 데이터 특징 항목

Candidate Generation 단계에서는 전처리된 데이터에 DeepFM 및 TF-IDF 를 병렬적으로 각각 적용하여 사용자의 취향이 반영된 콘텐츠 후보군을 선정하였다. DeepFM 학습 과정에서는 BCELoss 및 adam optimizer, Dropout rate 0.2, Embedding 차원을 100으로 설정하여 1400회 이상의 반복 학습을 수행하였다.

또한, 프로그램의 설명 스크립트, 프로그램명,

프로그램 ID 데이터 요소를 활용하여 데이터를 키워드화 하고, TF-IDF를 적용하여 코사인 유사도를 계산함으로써 콘텐츠 간 유사도를 산출하였다. TF-IDF를 통해 기존의 장르 기반의 추천 결과보다 콘텐츠 기반의 추천을 제공하여 사용자에게 다양한 추천이 가능하다는 장점이 있다. TF-IDF의 처리과정 및 연산 결과는 그림 2와 같다.

1. 사용 데이터 특징 (Feature)	2. 자연어처리 결과																											
<ul style="list-style-type: none"> 프로그램 (예)아이, 프로그램, 프로그램, 프로그램설명 장르: 프로그램 설명 키워드화하여 프로그램 간 유사도를 계산하여 유사한 프로그램들을 추천 	<p>[Input] CSI 25년 경력의 경찰 배대령 일 거리를 범행으로 과학수사대 요원들이 대학로 모퉁이 도시 리스캐니즘을 뒤집는 가장 범죄 현장을 다스려 과학적인 방법으로 증거를 채취하고 사건을 해결하는 드라마.</p> <p>[결과] 키워드 산출 최상위 과학 기술과 전통적인 경찰 수사 방식의 조화로써 사건을 해결해 나가는 내이-메이드 키언의 경찰 소수 과학수사대의 범죄를 신박한 액션으로 풀어내며 드라마.</p>																											
<p>3. TF-IDF 연산 결과</p> <table border="1"> <thead> <tr> <th>용량/언어 Index</th> <th>코사인 유사도</th> <th>* 문장/언어 Index</th> </tr> </thead> <tbody> <tr> <td>(0, 9462)</td> <td>0.8674019191740709</td> <td>(0) 프로그램: CSI</td> </tr> <tr> <td>(0, 17189)</td> <td>0.8234409895084602</td> <td>(0, 9469): 경찰</td> </tr> <tr> <td>(3, 3837)</td> <td>0.1592569189496992</td> <td>(0, 17189): 과학수사대</td> </tr> <tr> <td>(3, 19494)</td> <td>0.1199145400070304</td> <td></td> </tr> <tr> <td>(3, 7638)</td> <td>0.13771362632644538</td> <td></td> </tr> <tr> <td>(3, 15489)</td> <td>0.28247344131093044</td> <td></td> </tr> <tr> <td>(3, 14867)</td> <td>0.2102962988916145</td> <td></td> </tr> <tr> <td>(3, 8392)</td> <td>0.1839772989538994</td> <td></td> </tr> </tbody> </table>	용량/언어 Index	코사인 유사도	* 문장/언어 Index	(0, 9462)	0.8674019191740709	(0) 프로그램: CSI	(0, 17189)	0.8234409895084602	(0, 9469): 경찰	(3, 3837)	0.1592569189496992	(0, 17189): 과학수사대	(3, 19494)	0.1199145400070304		(3, 7638)	0.13771362632644538		(3, 15489)	0.28247344131093044		(3, 14867)	0.2102962988916145		(3, 8392)	0.1839772989538994		
용량/언어 Index	코사인 유사도	* 문장/언어 Index																										
(0, 9462)	0.8674019191740709	(0) 프로그램: CSI																										
(0, 17189)	0.8234409895084602	(0, 9469): 경찰																										
(3, 3837)	0.1592569189496992	(0, 17189): 과학수사대																										
(3, 19494)	0.1199145400070304																											
(3, 7638)	0.13771362632644538																											
(3, 15489)	0.28247344131093044																											
(3, 14867)	0.2102962988916145																											
(3, 8392)	0.1839772989538994																											

그림 2. TF-IDF 처리과정 및 연산 결과

랭킹 단계에서는 개인화된 콘텐츠 랭킹을 바탕으로 고객에게 맞춤형 콘텐츠를 추천하기 위해, 후보군 선정 단계에서 도출된 결과 값과 고객의 시청 정보, 출연진, 감독, 스텝, 본방 시작 시간 정보를 추가적으로 결합하여 DeepFM의 입력으로 사용한다.

학습된 모델의 추천 성능 검증을 위해 20명의 피험자를 바탕으로 성능을 측정하였다. 피험자는 OTT 플랫폼 내 약 10만명의 사용자 중 콘텐츠

시청 기록이 2000위~3000위 정도의 상위권 사용자 중 랜덤으로 선발하여 테스트를 진행하였다. 전체 모델의 Epoch 마다 도출된 추천 모델을 적용하여 test dataset의 추천 정확도를 산출하였다.

4. 결과

본 논문에서 제안한 OTT 서비스 내 콘텐츠 추천 모델은 총 25 회 학습을 진행하였다. 학습에 따른 추천 성능 결과는 그림 3과 같다.

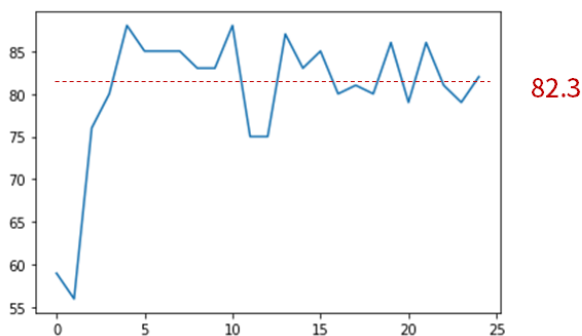


그림 3. 추천 모델 학습에 따른 추천 결과

추천 정확도는 추천 순위 1위에서 15위의 추천 콘텐츠를 시청자가 클릭하고 실제 시청이 이루어진 경우를 추천이 성공한 것으로 추정하여 정확도를 계산하였다. 실제 시청이 이루어진 경우는 전체 콘텐츠 길이의 80%이상을 시청한 경우를 실제 시청이 이루어졌다고 간주하였다. 만약, 추천된 콘텐츠를 사용자가 두 번 이상 시청한 경우에도 콘텐츠 추천을 성공한 것으로 판단하며, 시청 시간을 정규화 하여 한 번 시청한 것과 동일하게 결과를 계산하였다.

각 학습주기마다 학습된 모델을 바탕으로 테스트 데이터를 통해 성능 평가를 진행하였으며 추천의 정확도는 평균 82.3%이다. 약 5 Epoch까지 추천 성능이 급격하게 향상되었으며, 10 Epoch에서는 87%이상의 추천 성능을 보였다. 15 Epoch부터는 성능이 수렴하는 결과를 보였다.

4. 결론

본 논문에서는 OTT 서비스의 데이터를 기반으로 DeepFM 및 TF-IDF 알고리즘을 적용하여 개인화된 콘텐츠 추천 방법을 제안하였다. 제안된 방법은 피험자 20명을 대상으로 82%이상의 우수한 콘텐츠 추천 성능을 보였다. 그러나, 제안된 방법에서는 시리즈로 편성된 콘텐츠 위주의 추천이 이루어졌고, 개별 콘텐츠는 순위가 낮아져 추천되지 않는 경향을 보였다. 이에 향후, 본 논문에서 제안된 알고리즘을 기반으로 개별 콘텐츠 및 시리즈 콘텐츠에 가중치를 부여함으로써 콘텐츠 특징을 반영한 추천이 제공될 수 있도록 연구하고, A/B test 프레임워크를 도입하여 추천의 결과를 비교분석 할 예정이다.

[참고문헌]

- [1] Jieun Son, SeungBum Kim, Hyunjoong Kim, and Sungzoon Cho. "Review and Analysis of Recommender Systems." *Journal of the Korean Institute of Industrial Engineers*. 41.2(2015) pp. 185-208.
- [2] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191-198.
- [3] Gomez-Uribe, C., & Hunt, N. (2015). The Netflix Recommender System. *ACM Transactions on Management Information Systems (TMIS)*, 6, 1 - 19.
- [4] Ramos, J. (2003). "Using TF-IDF to Determine Word Relevance in Document Queries," *Proceedings of the First Instructional Conference on Machine Learning*, pp. 1-4
- [5] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. Product based neural networks for user response prediction. *CoRR*, abs/1611.00144, 2016.
- [6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishikesh Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. *CoRR*, abs/1606.07792, 2016.
- [7] Weinan Zhang, Tianming Du, and Jun Wang. Deep learning over multi-field categorical data - A case study on user response prediction. In *ECIR*, 2016.
- [8]. Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. In *IJCAI*. 2782-2788.