Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 3, pp. 280-288, June 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.3.280

<Application Research>

콘텐츠 선호 모형을 결합한 행렬 분해 기반 영화 추천시스템

백서인[1] · 민대기[2†]

[1]이화여자대학교 빅데이터분석학 협동과정 / [2]이화여자대학교 경영학과

Contents Preference Model Combined with Matrix Factorization for Movie Recommendation

Seoin Baek[1] · Daiki Min[2]

[1]Graduate School(Big Data Analytics), Ewha Womans University
[2]School of Business, Ewha Womans University

With the growth of the media market, companies that provide contents services such as movies, music and video are providing various content to satisfy users. While these changes have allowed users to enjoy richer content, a new problem has emerged that they have to spend much more time than before to find content that suits their taste among the overflowing content. Recommender system has become an important key to solve these problems. Matrix Factorization (MF) is the most well-known and widely used for identifying users' preference on contents. However, MF has a drawback of data sparsity and is not capable of utilizing meta-data. In this study, we proposed a two-stage contents preference model with Matrix Factorization (CPMF). The proposed method combines MF and contents preference models that utilize a variety of meta-data (e.g., actors, directors, and genres) to identify users' preferences. A numerical analysis is conducted to evaluate the performance of the proposed method for movie recommendation domains.

Keywords: Movie Recommender System, Matrix Factorization, Contents Preference Model

1. 서 론

1.1 연구 배경

OTT(Over-The-Top) 플랫폼 시장이 사용자 콘텐츠 소비의 중심으로 자리매김하며 데이터를 활용한 컨텐츠 추천 기술이 영화 추천을 중심으로 많이 사용되고 있다. 국내 시장조사 업체 메조미디어(http://www.mezzomedia.co.kr)의 조사결과에 따르면 2016년 4,884억 원 규모였던 국내 OTT 시장은 2019년 6,345억 원 규모로 성장했으며, 2020년에는 7,801억 원 규모가 될 것으로 전망하고 있다. 이에 따라 OTT 플랫폼 기업들은 사용자의 영화 콘텐츠 취향을 더욱 정교하게 파악하여 만족도를 극대화할 수 있는 추천 알고리즘 개발이 필수적인 요소가 되었다.

본 연구는 국내 한 IPTV 서비스 기업을 대상으로 컨텐츠 추천 모형을 개발하는 과정에서 분석한 사용자의 컨텐츠 선호 특성을 기반으로 한다. 해당 IPTV 서비스 기업에서 제공하는 서비스를 이용하는 사용자들의 VOD(Video On Demand) 시청 행태를 분석한 결과, 대부분의 사용자는 과거의 VOD 시청 행태와 매우 비슷한 방식으로 유사 VOD를 지속적으로 시청하는 것을 발견할 수 있었다. 약 일주일 기간 동안의 시청 이력 데이터를 분석한 결과를 몇 가지 흥미로운 점을 확인할 수 있었다. 첫째, 총 24개의 VOD 장르 중 상위 4개 장르(예능, 드라마, 시사교양, 액션영화)가 전체 시청 이력의 약 80%를 차지하였다. 사용자 한 명이 소비하는 컨텐츠의 장르는 매우 제한적으로, 약 2~4개 정도의 장르 내에서 컨텐츠를 시청하였다. 또한 동일 장르 내에서도 같은 시리즈의 영화, 드라마, 예능 등을 지속적으로 시청하는 것을 확인하였다. 이와 같은 사용자의 시청 행태를 살펴볼 때 장르, 출연진 등과 같은 메타데이터가

† 연락저자 : 민대기 교수, 03760 서울특별시 서대문구 이화여대길 52 이화여자대학교 경영학과, Tel : 02-3277-3923, Fax : 02-3277-2835, E-mail : dmin@ewha.ac.kr
2021년 1월 20일 접수; 2021년 3월 4일 수정본 접수; 2021년 3월 22일 게재 확정.

동영상 컨텐츠 추천에 있어 매우 핵심적인 요소임을 확인하였으며, 효과적인 메타데이터의 활용을 통하여 추천 알고리즘의 성능을 개선할 수 있는 가능성을 확인하였다. 연구의 배경이 되는 기업 데이터의 공개가 불가능한 상황을 고려하여, 본 연구에서는 기업 데이터 분석에서 확인한 메타데이터의 효용성을 공개 데이터를 사용하여 검증하도록 한다.

1.2 연구 목적

인터넷 기술의 발전 및 빅데이터 시대의 도래로 기업들은 사용자가 원하는 다양한 서비스를 제공하고 있다. 사용자는 여러 선택지 중에서 본인의 취향에 맞는 제품 혹은 서비스를 선택할 수 있게 되었지만, 너무 많은 선택지로 인해 취향에 맞는 제품을 찾기 위해 훨씬 많은 시간을 소비하고 있다. 이러한 이유로 사용자의 취향에 맞는 적절한 제품 혹은 서비스를 추천해 주는 것은 기업 입장에서 사용자의 만족도 및 충성도를 높일 수 있는 가장 좋은 방법 중 하나로 여겨져 왔다(Koren *et al*., 2009).

추천 알고리즘은 사용자를 충성 고객으로 만들어 서비스로 유입시키고 매출 증대를 가져올 수 있어 IT 서비스를 제공하는 기업에게는 매우 핵심적인 기술이라 할 수 있다. 특히 행렬 분해 알고리즘은 도메인에 대한 제약이 없고 사용자의 구매, 시청 등의 이력 데이터만 있으면 추천 알고리즘을 구축 가능하며 경험적으로 다른 알고리즘에 비해 우수한 성능을 보장한다는 점에서 추천 서비스를 도입하는 여러 분야에서 널리 사용되고 있다(Frolov and Oseledets, 2017).

추천시스템에 있어 행렬 분해(Matrix Factorization) 방법론의 장점은 평점(rating)과 같은 명시적 피드백(explicit feedback)이 존재하지 않을 때, 구매 이력, 검색 이력, 검색 패턴, 마우스 움직임 등과 같은 암시적 피드백(implicit feedback)을 사용하여 사용자의 선호도를 추론할 수 있다는 것이다(Symeonidis and Zioupos, 2016). 하지만 기존의 행렬 분해 방법론을 영화 추천 도메인에 적용할 때에는 다음과 같은 몇 가지 단점이 존재한다.

첫째, 사용자와 아이템 간의 상호작용만을 통해 아이템을 추천하게 되기 때문에 콘텐츠의 다양한 메타데이터를 활용하지 못한다는 것이다. 영화 추천 도메인에서는 아이템의 특성에 대해 파악할 수 있는 장르, 배우, 감독, 시놉시스 등 다양한 정보가 존재하기 때문에, 이러한 정보를 활용하면 추천 성능을 더욱 향상시킬 수 있을 것이다. 따라서, 최근에는 사용자와 아이템뿐만 아니라 새로운 변수를 추천시스템에 적용하려는 시도가 계속되고 있다(Frolov and Oseledets, 2017).

둘째, 데이터 희소성(Data sparsity) 문제이다. 이는 적게는 수백 개에서 많게는 수 십만 개의 아이템 중에 각 사용자가 실제 구매 혹은 시청한 아이템은 매우 소수이기 때문이다. 그렇기 때문에, 행렬의 대부분이 0으로 채워져 행렬이 매우 희소(sparse)하게 되고 이는 결국 제대로 된 학습을 하지 못하는 원인이 된다. 이러한 문제를 해결하기 위하여 사용자-아이템 행렬의 희소성을 줄이기 위해 신뢰할 수 있는 값으로 행렬을 채우거나 사용 가능한 추가적인 정보를 사용하여 행렬 분해 방법론을 확장하는 방법을 사용하고 있다(Parvin *et al*., 2019). 본 논문에서는 두 번째 방법과 같이 데이터 희소성에 의해서 추천 알고리즘의 성능이 낮은 것을 보완하기 위하여 추가 정보를 사용하는 방법을 제안하였다. 즉, 행렬 분해 방법론에 활용 가능한 추가적인 정보를 결합할 수 있는 새로운 방법론을 제안하고 데이터 희소성 문제를 해결하고자 한다.

본 논문은 다음과 같이 구성된다. 제 2장에서는 영화 도메인에서의 추천 시스템과 본 논문에서 제안하는 알고리즘의 기본 방법론인 행렬 분해 방법론의 선행 연구에 대해 설명한다. 제 3장에서는 2단계 콘텐츠 선호 모형을 결합한 CPMF(Contents Preference model combined with Matrix Factorization) 모델을 제안한다. CPMF 방법론은 다음과 같은 절차로 구성된다. 첫째, 행렬 분해 방법을 통해 사용자 별 각 아이템에 대한 점수를 계산한다. 둘째, 과거 시청 이력을 토대로 사용자의 배우, 감독 선호도를 계산 후 이를 가중치 함수를 통해 사용자가 선호하는 배우, 감독이 참여한 영화에 가중치를 부여한다. 셋째, 최종적으로 상위 n개의 영화를 추천할 때, 사용자의 장르 선호도를 통해 사용자의 선호도가 높은 장르를 대상으로 추천 리스트를 구성한다. 제 4장에서는 무비렌즈 100K 데이터를 활용하여 본 논문에서 제안한 방법론과 기본 행렬 분해 방법론과의 성능 비교 실험 결과를 제시한다. 마지막으로 제 5장에서는 논문의 의의와 한계를 서술하였다.

2. 선행연구

2.1 영화 추천 시스템

추천 알고리즘은 사용자의 행태를 분석하여 사용자가 선호할 것으로 예상되는 아이템을 추천한다는 점에서 고객 만족도 향상에 크게 기여해왔다. 이러한 이유로 다양한 도메인에서 추천 알고리즘을 적용한 서비스를 제공하고 있다. 영화 도메인 역시 추천 알고리즘이 널리 적용되고 있으며, 영화 추천 알고리즘에 대한 많은 연구가 진행되고 있다. 특히, 영화 장르뿐만 아니라 사용자들이 남긴 태그, 리뷰 등의 다양한 데이터를 활용한 연구가 활발히 진행되어 왔다(<Table 1> 참조).

Table 1. Movie Recommender System Using Meta-Data

Meta data	Reference
Genre	Choi *et al*.(2012), Reddy *et al*.(2019), Deldjoo *et al*.(2019), Su *et al*.(2020)
Review	Diao *et al*.(2014), Hyeon *et al*.(2019)
Tag	Zhang *et al*.(2010), Stanescu *et al*.(2013), Wei *et al*.(2016), Deldjoo *et al*.(2019)
Actor	Chen *et al*.(2017)
Etc.	Chen *et al*.(2017), Deldjoo *et al*.(2019), Khalaji *et al*.(2020), Singla *et al*.(2020)

대부분의 연구가 메타데이터를 활용하여 아이템간 유사도를 계산하여 추천하는 방법에 기반하고 있었다. Choi *et al*.(2012)는 사용자에 대한 정보가 충분하지 않은 상황에서 영화 장르의 상관관계를 활용하여 기존 협업 필터링 방법론의 성능 개선이 가능함을 제시하였다. Reddy *et al*.(2019)은 사용자가 좋아할 것으로 예측되는 영화 장르의 상관관계에 기반한 콘텐츠 기반 필터링 방법론을 제안하였는데, 장르 행렬(genre matrix)과 평점 행렬(rating matrix)의 내적 결과를 대상으로 사용자 간 유클리디안 거리를 계산하여 추천 대상 영화를 결정하였다. Chen *et al*.(2017)는 콘텐츠 정보를 사용하여 뉴럴 네트워크를 학습시킨 후 이를 통해 영화 간 유사도를 계산하여 추천하는 방법을 제안하였으며, Deldjoo *et al*.(2019)는 Movie Genome이라 부르는 영화의 콘텐츠 정보를 활용한 아이템 기반 최근접 이웃(item-based nearest neighbor approach) 추천 알고리즘을 제안하였다. Khalaji *et al*.(2020)는 Tanimoto Reliability Similarity Measure(TRSM)라는 새로운 유사도 측정 방법을 제안하여 cold-start 문제를 해소하는 연구를 진행하였으며, Singla *et al*.(2020)는 Doc2vec과 tf-idf를 결합한 하이브리드 방법을 제안하고, 영화의 줄거리, 평점, 제작 국가, 제작 연도를 변수로 사용하여 영화간 유사도를 계산하였다. Su *et al*.(2020)는 유저간 유사도를 항상 일정한 스칼라 값으로 계산되는 것이 아닌 벡터 값에 기반하여 계산되도록 하여 추천하고자 하는 아이템의 특징에 따라 유저의 선호도를 다각도로 모델링하였다.

이외에도 영화 평점과 같은 정량적 정보와 함께 사용자 리뷰를 토픽 모델링, 감성 분석 기법 등을 활용해 협업 필터링에 반영하는 연구(Diao *et al*., 2014; Hyeon *et al*., 2019)와 태그(Tag)가 개인의 선호와 아이템 콘텐츠에 대한 정보를 담고 있다는 가정에 기반하여 user-item-tag의 관계(relation) 그래프를 활용한 연구가 진행되었다(Zhang *et al*., 2010; Stanescu *et al*., 2013; Wei *et al*., 2016).

본 논문에서 제안하는 방법론은 메타 데이터를 사용하여 사용자의 선호를 계산하는 방법에서 기존 연구와 차별성이 있다. 대부분의 연구가 장르, 태그 등의 데이터를 활용해 유클리디안 거리, 피어슨 상관관계 등의 유사도 측정 기법을 통해 사용자에게 아이템을 추천하는 방식에 기반하고 있다. 본 연구에서 제안하는 방법론은 유사도 측정 기법이 아닌 메타데이터에 기반한 가중치 함수를 통해 사용자의 선호를 추정함으로써 기존 연구와 비교하여 계산량을 줄이고 빠르게 적용할 수 있는 장점이 존재한다.

2.2 행렬 분해 기반 추천시스템

행렬 분해 기반 방법론은 협업 필터링 기반 추천시스템 중 가장 성공적인 모델로 평가받는 방법론이다. 행렬 분해 방법론은 사용자와 아이템의 잠재된 특징(Latent features)을 학습하여 본 행렬의 결측치를 추정함으로써 데이터 희소성 문제를 해결하고 추천 성능을 개선하게 된다. 가장 대표적인 행렬 분해 방법론으로 특이값 분해(Singular Value Decomposition, SVD) 방법론이 있으며, 이를 활용한 다양한 연구가 진행되어 왔다(Baltrunas *et al*., 2011; Hernando *et al*., 2016; Guan *et al*., 2017; Chen *et al*., 2018; Parvin *et al*., 2019).

행렬 분해 방법론을 이용한 추천시스템 연구에서는 데이터 희소성 문제를 해결하기 위하여 명시적 피드백과 암시적 피드백(implicit feedback)을 행렬 분해 방법론에 함께 사용하는 방법(Xue, *et al*., 2017; Chen *et al*., 2018) 인기 아이템에 대한 선호도가 더 높은 특성을 이용하여 예측 평점을 보정하는 방법(Guan *et al*., 2017), 아이템이 소비되는 상황 문맥(contextual situation)을 고려하는 방법(Baltrunas *et al*., 2011) 등을 사용하였다. 또한 Parvin *et al*.(2019)은 데이터 희소성과 확장성 문제(Scalability issue)를 해결하기 위하여 사용자의 social trust information을 활용하는 방법을 제안하였다.

앞서 살펴본 바와 같이 행렬 분해 방법론에 기반한 다양한 추천 방법론이 제시되었다. 본 논문에서도 데이터 희소성 문제를 해결하고 추천 성능을 개선하기 위하여 행렬 분해 방법론을 활용하도록 한다. 특히, 활용 가능한 메타데이터를 행렬 분해 방법론과 결합하여 추천 성능을 개선하도록 한다.

3. 제안 모형

본 논문에서 제안하는 방법론은 협업 필터링에 기반한 행렬 분해 방법론에 콘텐츠 기반 필터링에 사용되는 아이템에 대한 메타데이터 정보(배우, 감독, 장르)를 결합한 하이브리드 추천 기법(CPMF)을 제안한다. 먼저, 행렬 분해 방법론을 사용하여 사용자의 각 아이템에 대한 선호도 점수를 산출하고, 이를 대상으로 사용자의 배우, 감독, 장르 선호도를 반영하여 보정된 선호도 점수를 계산한다. 보정된 선호도 점수가 높은 순서대로 상위 n개의 아이템(Top-n items)을 추천한다. 마지막으로 추천 모형의 성능은 Recall@n 지표를 이용하여 측정하였다.

3.1 행렬 분해 방법

식 (1)은 대표적인 행렬 분해 방법인 특이값 분해(Singular Value Decomposition, SVD) 방법을 나타낸다.

$$A_{n\times m} \approx U_{n\times k} \cdot S_{k\times k} \cdot V^{T}_{k\times m} = \hat{A}_{n\times m} \tag{1}$$

$\hat{A}_{n\times m}$는 관측 데이터 A에 대한 예측 행렬(Prediction matrix)이며, U는 AA^{T}를 고유값 분해(Eigen value decomposition)해서 얻어진 직교 행렬(orthogonal matrix)이 된다. A의 좌측 특이벡터(left singular vector) V는 $A^{T}A$를 고유값 분해(Eigen value decomposition)해서 얻어진 직교 행렬(Orthogonal matrix)이며, A의 우측 특이 벡터(right singular vector) S는 AA^{T}와 $A^{T}A$를 고유값 분해하여 도출한 고유값들의 제곱근을 대각 원소로 하는

대각 행렬로 대각 원소들이 A의 특이값(Singular values)이 된다. 여기서 특이값 k는 하이퍼 파라미터이며, 이는 원래의 행렬 A에서 얼만큼의 정보를 보존할지를 결정한다(Symeonidis and Zioupos, 2016).

식 (1)을 식 (2)와 같이 재정의 할 수 있다.

$$A = UV^T, \text{where } V^T = S_{k\times k} \cdot V^T_{k\times m} \quad (2)$$

영화 추천 시스템을 대상으로 식 (1)과 식 (2)에서 제시한 행렬 분해 방법론을 설명하면 다음과 같다. 각 사용자가 아이템에 부여한 평점으로 사용자-아이템 행렬 $A_{n\times m}$를 만든다. 이후 사용자에 대한 잠재 특성 행렬 U, 아이템에 대한 잠재 특성 행렬 V를 임의의 값으로 초기화 한다. 행렬 분해 알고리즘을 통해 각 사용자의 잠재 벡터와 아이템의 잠재 벡터의 내적을 구하여 사용자의 아이템 별 평점을 예측한다($\hat{A}_{n\times m}$). 이후에 예측 값과 실제 값의 오차를 최소화하는 방향으로 행렬 U와 V를 반복적으로 학습하며, 학습된 행렬 U와 V로부터 사용자가 아직 평가하지 않은 아이템에 대한 예상 평점을 예측할 수 있다. 본 논문에서는 행렬 U와 V로부터 도출한 예측 값을 콘텐츠 선호 모형을 통해 보정하여 사용자의 선호도를 파악하는 방법론을 제안한다.

3.2 배우 & 감독 선호도

본 논문에서는 사용자가 선호하는 배우와 감독이 있다면, 해당 배우가 출연하거나 해당 감독이 연출한 영화는 높은 확률로 시청할 것이라는 연구가설에 기반하여 방법론을 제안하였다. 따라서 학습 데이터셋에서 각 사용자가 시청했던 영화에 출연한 배우와 영화를 연출한 감독에 대한 데이터를 추가로 수집하고 이를 이용하여 사용하여 행렬 분해 기반 방법론으로부터 나온 예측 값을 보정하였다. 즉, 행렬 분해로부터 나온 선호도 예측값에 사용자가 각 영화에 출연한 배우와 감독에 얼만큼의 선호를 갖고 있는지 추가적인 정보를 사용하여 가중치를 부여하였다.

배우와 감독 선호도는 학습 데이터셋에서 각 배우와 감독이 나온 영화를 시청한 횟수를 사용하였다. 이때, 배우에 대한 선호도를 구하기 위해 해당 영화에 출연한 주연 배우 2명에 대해 각 주연 배우의 다른 작품을 몇 개 시청한 이력이 있는지를 계산하였고, 감독 선호도는 해당 영화를 연출한 감독 1명에 대해 다른 작품을 몇 개 시청한 이력이 있는지를 계산하였다. 이 횟수의 총합을 시그모이드 함수(sigmoid function)에 적용하여 계산한 값을 가중치로 사용하였다.

배우 & 감독 선호도에 대한 예시는 <Figure 2>와 같다. 사용자 i의 영화 j에 대한 점수를 구하기 위해 영화 j의 콘텐츠 정보를 활용한다. 영화 j는 배우 B와 C가 출연하며, 감독은 a 이다. 사용자 i는 배우 B가 나온 영화는 1개, 배우 C가 나온 영화 4개, 감독 a가 연출한 영화를 1개 시청한 이력이 있다. 이 횟수의 총합 6이 사용자 i의 영화 j에 대한 배우 & 감독 선호도가 된다. 이 횟수를 가중치 함수(weight function)에 넣어 최종 가중치 값을 구한다. 최종 가중치 값은 식 (3)과 같다. 이와 같이 구한 값을 행렬 분해로부터 나온 예측치 $\hat{r}_{ij}$에 곱하여 최종 점수(Final score)를 구한다.

$$Weight = 1 + f(n) \quad (3)$$

3.3 장르 선호도

본 논문은 각 사용자는 자신이 선호하는 장르가 있으며, 이는 향후 영화 선택에 유의미한 영향을 미친다는 연구가설에 기반하고 있다. 즉, 코미디, 액션 영화를 주로 시청한 이력이 있는 사용자는 앞으로도 코미디 혹은 액션 영화를 시청할 확률이 높다고 가정한다.

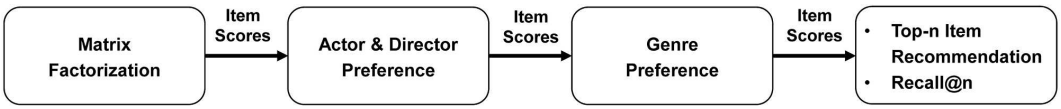

Figure 1. Proposed Model-Contents Preference Model Combined with Matrix Factorization

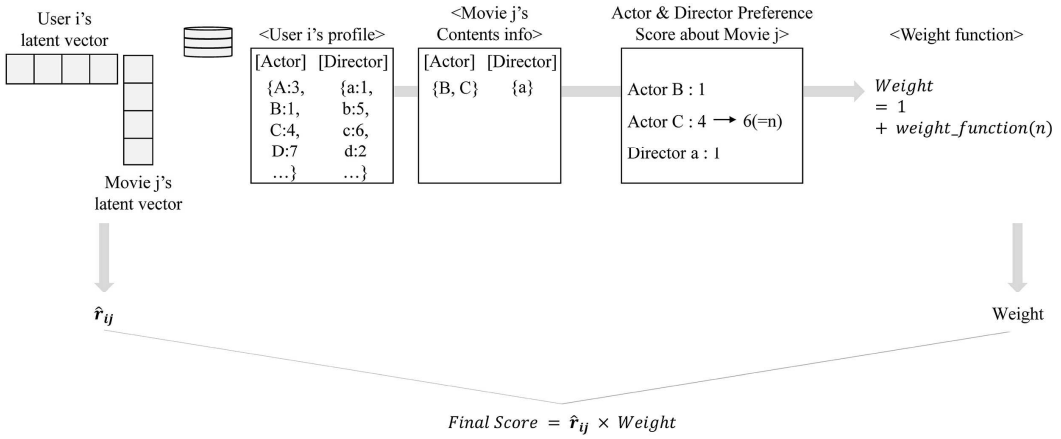

Figure 2. Actor and Director Preference

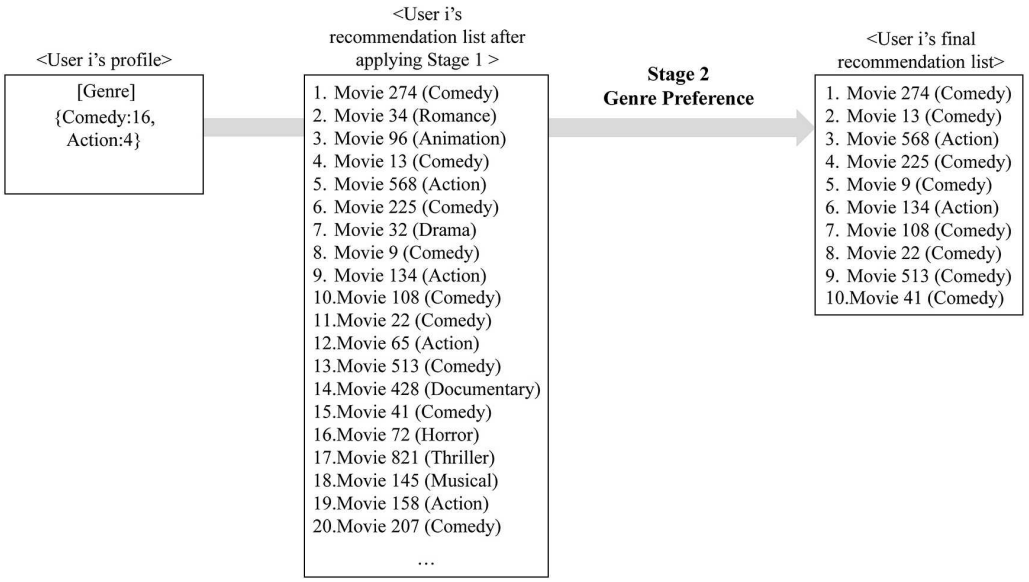

Figure 3. Genre Preference

<Figure 1>에서 제시한 바와 같이 행렬 분해 값으로 학습된 점수에 배우 & 감독 선호도를 반영한 값이 사용자가 각 영화에 대하여 갖고 있는 선호도 점수가 되며, 이 점수가 높은 순서대로 영화를 사용자에게 추천하게 된다. 최종적으로 추천할 n개의 아이템 (즉, 영화) 리스트 구성할 때, 각 사용자가 기존에 시청하였던 장르의 비율을 반영하여 추천 리스트를 구성한다. 예를 들어, 사용자의 시청 이력 중 코미디 장르가 80%를 차지하고, 액션 장르가 20%를 차지한다면, 추천할 영화 리스트도 이 비율에 맞춰 코미디 영화가 80%, 액션 영화가 20%를 차지하도록 구성한다. <Figure 3>은 장르 선호를 반영하는 방법의 예를 나타낸다.

장르 선호도를 반영하는 것은 다음과 같은 장점을 갖는다. 사용자의 선호와 전혀 맞지 않는 장르의 영화가 추천 리스트에 올라왔을 때, 이를 필터링하는 것이 가능하다. 일반적으로 협업 필터링 방법은 사용자의 선호와 상관없이 여러 사람들에게 인기있는 아이템이 추천 리스트에 지속적으로 포함되는 특징이 있다. 따라서, 장르 선호도를 명시적으로 반영함으로써 사용자 선호와 관련 없는 영화가 추천리스트에 포함되는 문제를 어느정도 개선하는 것이 가능하다.

3.4 성능 평가

기존의 많은 연구에서는 추천시스템의 성능을 평가하기 위해 RMSE, MAE 등의 오차 평가지표를 사용하였는데, 이와 같은 지표는 추천을 평점 예측(Rating prediction)의 문제로 보고 영화 평점 예측의 정확도를 평가하기 위한 목적에 적합하다. 하지만 현실적으로 추천시스템의 역할은 사용자의 평점을 예측하는 것보다 사용자가 실제 구매하거나 또는 시청할 만한 영화 아이템을 찾는 것에 가깝다. 이 경우 선호도 예측 결과를 기반으로 상위 n개의 영화 아이템을 추천 대상으로 결정하고 실제 사용자의 선택 여부를 평가하는 것이 필요하다(Top-n 아이템 추천 방식). Top-n 아이템 추천 방식의 경우 전통적인 오차 평가지표는 적합하지 않으며, Recall과 같은 정확도 측정지표(Accuracy metrics)가 Top-n 추천 방식의 성능 평가 방식으로 사용되고 있다. 본 논문에서도 제안하는 추천 방법론의 성능 평가를 위해 Recall@n을 사용하였다.

본 논문에서 제안 방법론의 성능을 평하기 위하여 사용한 Recall@n은 식 (4)를 이용하여 계산한다. Relevant items는 사용자가 실제 시청한 아이템을 의미하며, recommended items는 추천 알고리즘이 추천한 아이템을 의미한다. 예를 들어, recall@30은 다음과 같이 계산할 수 있다. n = 30이고 사용자가 시청한 10개의 아이템 중 2개의 아이템이 추천 알고리즘에 의해 추천된 아이템이라면 recall@30 = 2/10 = 0.2가 된다.

$$Recall@n = \frac{relevant\ items \cap recommended\ items}{relevant\ items} \quad (4)$$

4. 수치실험

4.1 자료 수집

본 논문에서는 무비렌즈 100K 데이터셋(https://grouplens.org/datasets/movielens)을 이용하여 제안 방법론의 성능을 평가하였다. 무비렌즈 100K 데이터셋은 사용자 943명, 영화 1,642편, 총 100,000개의 영화 리뷰 평점으로 이루어진 데이터셋으로 사용자 별로 약 20개 이상의 영화 리뷰 평점(Rating) 데이터를 포함하고 있다. 사용자의 배우, 감독에 대한 선호도를 파악하기 위해 각 영화의 배우, 감독에 대한 정보를 인터넷 영화 데이터베이스(https://www.imdb.com)를 통해 추가로 수집하였다.

4.2 데이터 탐색적 분석

본 논문에서 제안하는 방법론은 사용자의 선호를 예측하는 데 도움이 될 수 있는 다양한 메타 데이터(배우, 감독, 장르)를 행렬 분해 방법론과 결합하는 것이다. 이는 사용자가 기존에 시청하였던 영화들이 가진 공통적인 특성(배우, 감독, 장르)이 있다면, 이후에도 동일 특성을 가진 다른 영화를 시청할 확률이 높다는 가정에 기반한다. 이를 확인하기 위하여 데이터 탐색적 분석을 먼저 수행하였다.

(1) 배우 선호도

사용자의 배우 선호도를 확인하기 위해, 각 사용자가 적어도 3편 이상 시청한 배우가 몇 명인지 확인하였다. 전체 사용자 943명중 65.32%에 해당하는 616명이 동일 배우의 작품 3개 이상을 시청한 이력이 있었으며, 616명의 사용자는 평균 15.13명의 배우에 대해 작품 3개 이상을 시청하였고, 최소값은 1명, 최대값은 114명이었다(<Table 2> 참조).

Table 2. Descriptive Statistics for the Number of Actors Who have been Chosen more than Three Movies by Each User

Users	Mean	Std	Min	25%	50%	75%	Max
616	15.13	18.42	1	2	8	21	114

(2) 감독 선호도

사용자의 감독 선호도를 확인하기 위해, 각 사용자가 시청한 영화 중에서 동일한 감독이 연출한 영화를 3편 이상 시청한 경우를 확인하였다. 전체 사용자 943명중 50.9%에 해당하는 480명이 동일 감독의 작품 3편 이상을 시청한 이력이 있었다. 480명의 사용자는 평균 7.42명의 감독에 대해 영화 3편 이상을 시청하였고, 최소값은 1명, 최대값은 42명이다(<Table 3> 참조).

Table 3. Descriptive Statistics for the Number of Directors who have been Chosen more than Three Movies by Each User

Users	Mean	Std	Min	25%	50%	75%	Max
480	7.42	7.24	1	2	5	10	42

(3) 장르 선호도

사용자의 장르 선호도를 확인하기 위해, 사용자가 시청한 전체 영화 수에서 사용자가 가장 많이 시청한 장르가 속한 영화 수의 비율을 알아보았다. 각 영화는 한 개 이상의 장르에 포함되어 있기 때문에 총 19개의 장르에 대해 사용자가 시청한 영화가 포함된 장르를 모두 시청 횟수에 반영하였다. 장르 시청 비율(각 사용자가 가장 많이 시청한 장르가 포함된 영화 수/각 사용자가 시청한 영화 총 개수)은 평균 48.16%였으며, 최소값은 29.52%, 최대값은 95.65%이다.

Table 4. Descriptive Statistics for the Percentage of the most watched Genre in the Entire Viewing List by Each User

Users	Mean	Std	Min	25%	50%	75%	Max
943	48.16%	11.25%	29.52%	40.00%	45.61%	53.57%	95.65%

사용자 별 두 번째, 세 번째로 가장 많이 시청한 장르가 전체 시청 리스트에서 차지하는 비율을 함께 분석하였다. 사용자가 두 번째로 가장 많이 시청한 장르의 비율은 평균 33.10%이며, 최소값과 최대값은 각각 16.28%와 61.90%였다. 사용자가 세 번째로 가장 많이 시청한 장르는 평균 26.64%이며, 최소값은 11.76%, 최대값은 54.17%이다.

Table 5. Descriptive Statistics for the Percentage of the Second most watched Genre in the Entire Viewing List by Each User

Users	Mean	Std	Min	25%	50%	75%	Max
943	33.10%	6.36%	16.28%	29.23%	32.24%	36.36%	61.90%

Table 6. Descriptive Statistics for the Percentage of the Third most watched Genre in the Entire Viewing List by Each User

Users	Mean	Std	Min	25%	50%	75%	Max
943	26.64	5.36	11.76	22.99	26.52	29.82	54.17

데이터 탐색적 분석을 통해 살펴본 결과, 전체 사용자의 65.32%가 동일 배우가 출연한 작품 3편 이상을 시청한 이력이 있었으며, 50.9%의 사용자는 동일 감독이 연출한 작품 3편 이상을 시청하였다. 이는 사용자가 특정 배우와 감독에 대해서 선호를 가지고 해당 배우 혹은 감독의 다른 작품도 시청하였다고 판단할 수 있다. 또한, 사용자는 본인이 선호하는 장르가 있으며, 해당 장르가 사용자의 시청 리스트에서 많은 비중을 차지하고 있음을 알 수 있었다. 이를 통해, 사용자의 배우, 감독 선호도, 그리고 장르 선호도를 반영한다면 사용자의 선호도를 더욱 잘 파악하는 것이 가능할 것으로 예상할 수 있다.

4.3 실험 결과

본 논문에서 제안하는 방법론(CPMF)의 성능을 확인하기 위해 다음과 같이 3개의 추가 방법론을 함께 고려하여 실험을 진행하였다.

- MF_base : 행렬 분해 방법론 만을 적용하여 도출한 선호도 점수를 기반으로 추천 리스트를 결정한 경우
- MF_actor : 행렬 분해 방법론에서 도출한 선호도 점수에 배우와 감독 선호도의 가중치를 부여하여 선호도 점수를 보정한 경우
- MF_genre : 행렬 분해 방법론에서 도출한 선호도 점수를 기준으로 추천리스트를 결정하되 장르 선호도를 고려하여 추천 대상을 필터링한 경우

• CPMF : 본 논문에서 제안하는 방법론으로 행렬 분해 방법론에서 도출한 선호도 점수에 배우와 감독 선호도의 가중치를 부여하여 선호도 점수를 보정하고, 마지막으로 장르 선호도를 적용하여 추천 대상 영화를 필터링한 경우

실험을 위한 기본 환경은 <Table 7>과 같다. 추천 개수 n 30개, 과적합 방지를 위한 정규화 파라미터 0.1, 학습률 0.005, 행렬 분해 방법론의 잠재 요인 20, 배우와 감독 선호도 가중치 함수는 시그모이드 함수를 사용하였다. 또한 제안 방법론의 성능평가를 위하여 5차 교차검증(5-fold cross validation)을 수행하였다.

Table 7. Experimental Environment

n	Epoch	Regularized factor	Learning rate	Latent factor	Weight function
30	500	0.1	0.005	20	sigmoid

(1) 제안방법론의 성능 평가

실험 결과는 <Figure 4>와 같다. 본 논문에서 제안하는 CPMF 방법론의 Recall@30이 0.1235로 가장 높은 성능을 보였으며, MF_actor, MF_genre, MF_base 방법론이 각 0.1057, 0.0669, 0.0329의 성능을 보였다. 이를 통해 본 논문에서 제안하는 CPMF 방법론이 행렬 분해 방법론의 성능 향상에 도움이 되는 것을 확인할 수 있었다. 또한, 사용자의 배우와 감독 선호도를 고려한 MF_actor 방법론이 사용자의 장르 선호도를 반영한 MF_genre 방법론에 비해 좋은 성능을 보였는데, 이는 장르의 경우 동일 장르라 하더라도 수많은 영화가 존재하기 때문에, 사용자의 개인화된 선호를 파악하기에 어려운 반면, 배우와 감독 선호도는 사용자의 실제 시청 이력을 바탕으로 하기 때문에 장르 선호도에 비해 사용자의 개인화된 선호를 잘 파악한 것으로 해석할 수 있다. 하지만 MF_actor 방법론과 MF_genre 방법론 모두 기존 행렬 분해 방법론인 MF_base와 비교해 성능 향상이 있었으며, 배우와 감독 선호도, 장르 선호도를 함께 고려한 CPMF 방법론이 가장 큰 성능 향상을 보임을 알 수 있었다.

추천 개수를 5, 10, 20, 30으로 바꾸어가며 실험을 진행하였다. 실험 결과, 본 논문에서 제안하는 CPMF 방법론이 모든 경우에서 가장 우수한 성능을 보였으며, MF_actor, MF_genre, MF_base 방법론 순으로 높은 성능을 보였다. 이를 통해 CPMF 방법론이 사용자의 선호를 파악하여 관련 아이템을 추천 리스트 상위에 위치시킬 뿐만 아니라 추천 개수가 늘어나도 사용자의 선호와 밀접한 아이템을 지속적으로 추천 리스트에 포함시키고 있음을 확인할 수 있다.

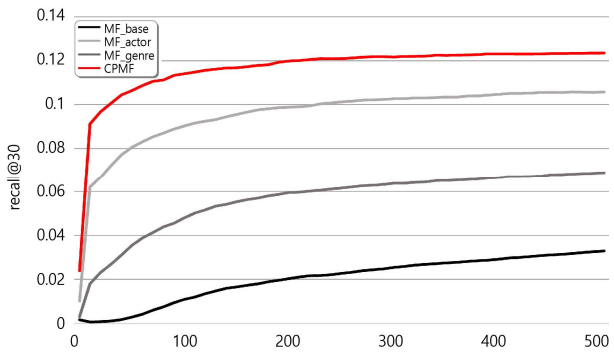

Figure 4. CPMF Recall@30

(2) 행렬 분해 방법론의 효과

MF_actor와 CPMF 방법론은 사용자의 시청 이력에 기반하여 사용자가 다수의 작품을 시청한 배우와 감독의 영화에 가중치를 주는 방식이다. 즉, 행렬 분해를 적용 후 나온 예측 값에 배우와 감독 선호도로 가중치를 주어 값을 보정한다. 행렬 분해 방법론을 사용하지 않고 단순히 선호도가 높은 배우와 감독의 영화를 추천하는 방법론과의 성능을 비교하기 위해 MF_actor 방법론과 CPMF 방법론에서 행렬 분해를 수행하지 않고 배우와 감독 선호도만 고려하여 성능을 측정해 보았다. 이에 대한 결과는 <Figure 5>와 같다. 실험 결과, MF_actor 방법론과 CPMF 방법론에서 모두 행렬 분해를 적용 후 배우와 감독 선호도로 가중치를 주었을 때 더욱 좋은 성능을 보였다. 이와 같은 결과는 행렬 분해 후 배우와 감독 선호도를 고려했을 때, 단순히 배우와 감독 선호도로 추천을 하는 것보다 사용자의 선호를 더욱 정교하게 파악할 수 있음을 의미한다.

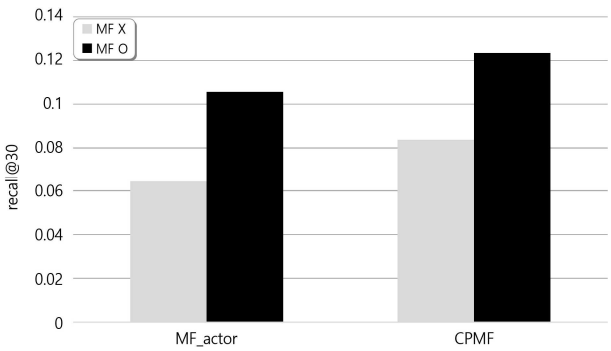

Figure 5. Recall@30 According to Applying Matrix Factorization

(3) 선호도 보정을 위한 가중치 함수

MF_actor 방법론과 CPMF 방법론에서는 식 (5)를 사용하여 배우와 감독 선호도에 가중치를 부여하였다. 식 (5)에서 $f(n)$은 가중치 부여 함수를 의미하며 기본 모형으로는 시그모이드 함수를 사용하였는데, 이와 함께 로그 함수($f(n)=\log_{10}^{n}$), 지수 함수($f(n)=e^{n}$), 선형 함수($f(n)=n$)를 사용하여 성능을 비교해 보았다.

$$Weight = 1 + f(n) \tag{5}$$

MF_actor 방법론과 CPMF 방법론에서 모두 시그모이드 함수로 가중치를 주었을 때 가장 좋은 성능을 보였으며, 로그 함수, 선형 함수, 지수 함수 순으로 좋은 성능을 보였다. 시그모이드 함수는 값이 0~1사이 값으로 나오기 때문에 배우와 감독 선호도가 아무리 높더라도 가중치의 최대값은 2(= 1+1)이다. 따라서 행렬

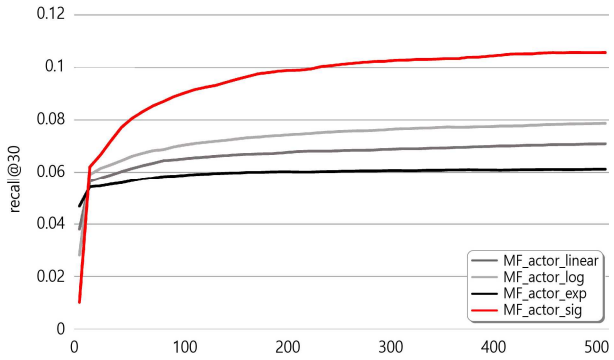

Figure 6. Recall@30 for MF_actor Model According to weight Function

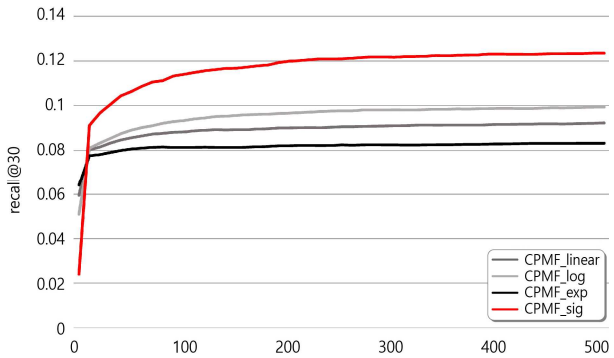

Figure 7. Recall@30 for CPMF Model According to weight Function

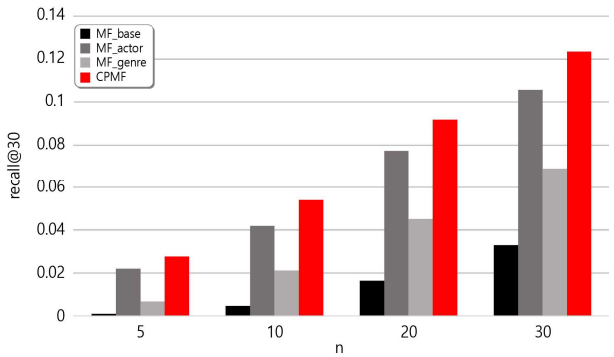

Figure 8. Recall@n According to the Number of Recommendation

분해를 통해 나온 값이 높지만, 사용자의 배우와 감독 선호도는 높지 않은 영화의 경우에도 추천 리스트에 포함될 수 있다. 반면, 지수 함수의 경우 배우와 감독 선호도가 높을수록 가중치의 값이 기하급수적으로 증가한다. 따라서 행렬 분해를 통해 나온 값이 높지만, 배우와 감독 선호도가 낮을 경우 추천 리스트에 포함되기 어렵다. 이는 결국 사용자의 잠재된 선호를 반영하지 못하고 추천 리스트를 다양하게 구성하지 못하도록 한다.

5. 결론 및 논의

본 연구는 사용자의 선호를 예측하는데 도움이 될 수 있는 다양한 메타 데이터(배우, 감독, 장르)를 사용한 2단계 콘텐츠 선호 모형을 행렬 분해 방법론과 결합한 CPMF(two-stage Contents Preference model combined with Matrix Factorization) 방법론을 제안하였다. 이는 사용자가 기존에 시청하였던 영화들이 가진 공통적인 특성(배우, 감독, 장르)이 있다면, 이후에도 동일 특성을 가진 다른 영화를 시청할 확률이 높다는 연구가설을 기반으로 하고 있다. 본 논문에서는 수치실험을 통하여 본 논문에서 제안한 CPMF 방법론이 기본 행렬 분해 방법론을 포함하는 다른 방법론과 비교하여 우수한 성능을 보임을 확인하였다.

본 연구는 다음과 같은 의의를 갖는다. 첫째, 메타데이터를 활용하여 아이템간 유사도를 계산하여 추천하는 기존 연구와 달리 본 연구에서 제안하는 방법론은 메타데이터에 기반한 가중치 함수를 통해 사용자의 선호를 추정하는 방법을 제안하였다. 둘째, 행렬 분해 방법론의 문제점으로 제시되는 데이터 희소성 문제를 영화 추천 도메인에서 다양한 메타데이터를 활용한 방법을 통해 개선하고자 하였다. 즉, 사용 가능한 추가적인 정보(배우, 감독, 장르 선호도)를 사용하여 행렬 분해 방법론을 확장하여 데이터 희소성 문제를 개선하였는데, 사용자의 배우와 감독 선호도를 수치화하여 가중치 함수를 통해 점수를 보정하였고, 장르 선호도로 필터링을 하여 사용자의 선호를 더욱 정교하게 파악하였다. 셋째, 협업 필터링과 콘텐츠 기반 필터링을 결합한 하이브리드 추천 방법을 제안하였다. 협업 필터링에 기반한 행렬 분해 방법론에 콘텐츠 기반 필터링에 사용되는 아이템에 대한 메타데이터 정보(배우, 감독, 장르)를 결합한 하이브리드 추천 기법을 제안하였다.

본 논문은 다음과 같은 한계점을 갖고 있다. 첫째, 본 논문에서 제안한 방법론은 recall@30 = 0.1235의 성능을 보였다. 이 결과는 본 논문에서 사용한 무비렌즈 100K를 대상으로 행렬분해 기반의 알고리즘을 제안한 타 논문에서 제시한 성능 값과 비교하여 우수하였다. 예를 들어, Slokom *et al.*(2020)은 Recall@5의 성능을 0.01~0.07로 제시하였으며, Polatidis and Georgiadis(2016)에서 제안하는 방법론의 Recall@10 값은 0.07 수준이었다. 이와 유사하게 Dong *et al.*(2017)의 연구에서도 Recall@50의 성능을 0.1~0.2로 제시하였다. 하지만, 컨텐츠 추천을 포함하여 다양한 분야에서 딥러닝(deep learning) 방법론을 이용한 연구가 제시되고 있으며, 성능 개선의 가능성을 보여주고 있다. 따라서 본 연구에서 확인한 메타 데이터의 활용 효용성을 딥러닝 기반의 방법론을 대상으로 검증하는 것을 고려할 수 있다.

둘째, 본 논문에서 제안하는 방법론의 제약은 한 명의 감독 혹은 배우가 참여한 영화의 수가 많지 않을 수 있으며, 이 경우 데이터 희소성 문제의 개선이 어렵다는 점이다. 따라서 동일 성향을 가진 감독 혹은 배우를 클러스터링 등의 방법을 통해 그룹화하여 사용자의 선호가 높은 클러스터에 해당하는 감독과 배우에게 가중치를 주어 추천하는 방법으로 향후 연구를 진행해 볼 수 있을 것이다. 해당 방법을 적용한다면 데이터 희소성 문제를 보다 근본적으로 개선하면서 사용자의 선호에 맞는 훨씬 다양한 아이템 추천이 가능할 것으로 생각한다.

마지막으로, 본 논문에서 제안한 방법론은 영화 추천 도메인을 대상으로 한 것으로 다른 도메인의 추천을 위해 적용하기는 어려움이 있다. 다른 도메인에 적용하기 위해서는 영화 추천에서 사용하였던 배우, 감독, 장르와 같이 아이템의 특성을 잘 파악할 수 있는 새로운 변수를 찾고 이를 검증하기 위한 추가적인 실험이 필요할 것으로 생각된다.

참고문헌

Chen, H. W., Wu, Y. L., Hor, M. K., and Tang, C. Y. (2017), Fully Content-based Movie Recommender System with Feature Extraction Using Neural Network, *In 2017 International Conference on Machine Learning and Cybernetics(ICMLC)*, **2**, 504-509.

Chen, S. and Peng, Y. (2018), Matrix Factorization for Recommendation with Explicit and Implicit Feedback, *Knowledge-Based Systems*, **158**, 109-117.

Choi, S. M., Ko, S. K., and Han, Y. S. (2012), A Movie Recommendation Algorithm based on Genre Correlations, *Expert Systems with Applications*, **39**(9), 8079-8085.

Deldjoo, Y., Schedl, M., and Elahi, M. (2019), Movie Genome Recommender : A Novel Recommender System based on Multimedia Content, *In 2019 International Conference on Content-Based Multimedia Indexing(CBMI)*, 1-4.

Diao, Q., Qiu, M., Wu, C. Y., Smola, A. J., Jiang, J., and Wang, C. (2014), Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation(JMARS), *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 193-202.

Dong, X., Yu, L., Wu, Z., Sun, Y., Yuan, L., and Zhang, F. (2017), A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems, *In Proceedings of the AAAI Conference on Artificial Intelligence*, **31**(1).

Frolov, E. and Oseledets, I. (2017), Tensor Methods and Recommender Systems, *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, **7**(3), e1201.

Guan, X., Li, C. T., and Guan, Y. (2017), Matrix Factorization with Rating Completion : An Enhanced SVD Model for Collaborative Filtering Recommender Systems, *IEEE Access*, **5**, 27668-27678.

Hernando, A., Bobadilla, J., and Ortega, F. (2016), A Non Negative Matrix Factorization for Collaborative Filtering Recommender Systems based on a Bayesian Probabilistic Model, *Knowledge-Based Systems*, **97**, 188-202.

Hyeon, J. Y., Yu, S. I., and Lee, S. Y. (2019), A Study on the Improvement of Recommendation System by Combining Ratings and Review Text Sensitivity Analysis, *Journal of Intelligence and Information Systems*, **25**(1), 219-239.

Khalaji, M. (2020), TRSM-RS : A Movie Recommender System Based on Users' Gender and New Weighted Similarity Measure, *arXiv preprint arXiv:2011.05119*.

Koren, Y., Bell, R., and Volinsky, C. (2009), Matrix Factorization Techniques for Recommender Systems, *Computer*, **8**, 30-37.

Parvin, H., Moradi, P., Esmaeili, S., and Qader, N. N. (2019), A Scalable and Robust Trust-Based Nonnegative Matrix Factorization Recommender Using the Alternating Direction Method, *Knowledge-Based Systems*, **166**, 92-107.

Polatidis, N. and Georgiadis, C. K. (2016), A Multi-Level Collaborative Filtering Method that Improves Recommendations, *Expert Systems with Applications*, **48**, 100-110.

Reddy, S. R. S., Nalluri, S., Kunisetti, S., Ashok, S., and Venkatesh, B. (2019), Content-based Movie Recommendation System Using Genre Correlation, *Smart Intelligent Computing and Applications*, 391-397.

Singla, R., Gupta, S., Gupta, A., and Vishwakarma, D. K. (2020), FLEX : A Content Based Movie Recommender, *In 2020 International Conference for Emerging Technology(INCET)*, 1-4.

Slokom, M., Larson, M., and Hanjalic, A. (2020), Partially Synthetic Data for Recommender Systems: Prediction Performance and Preference Hiding, *arXiv preprint arXiv:2008.03797*.

Stanescu, A., Nagar, S., and Caragea, D. (2013), A Hybrid Recommender System : User Profiling from Keywords and Ratings, *In 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence(WI) and Intelligent Agent Technologies(IAT)*, **1**, 73-80.

Su, Z., Zheng, X., Ai, J., Shen, Y., and Zhang, X. (2020), Link Prediction in Recommender Systems based on Vector Similarity, *Physica A : Statistical Mechanics and its Applications*, **560**, 125154.

Symeonidis, P. and Zioupos, A. (2016), *Matrix and Tensor Factorization Techniques for Recommender Systems*, New York : Springer International Publishing.

Wei, S., Zheng, X., Chen, D., and Chen, C. (2016), A Hybrid Approach for Movie Recommendation Via Tags and Ratings, *Electronic Commerce Research and Applications*, **18**, 83-94.

Xue, H. J., Dai, X., Zhang, J., Huang, S., and Chen, J. (2017), Deep Matrix Factorization Models for Recommender Systems, *In IJCAI*, 3203-3209.

Zhang, Z. K., Zhou, T., and Zhang, Y. C. (2010), Personalized Recommendation Via Integrated Diffusion on User-Item-Tag Tripartite Graphs, *Physica A : Statistical Mechanics and its Applications*, **389**(1), 179-18.

저자소개

백서인 : 이화여자대학교 경영학과에서 2018년 학사학위를 취득하고 2021년 2월 동 대학원 빅데이터분석학 협동과정 석사학위를 취득하였다. 연구분야는 빅데이터분석, 추천시스템이다.

민대기 : 서울대학교 산업공학과에서1999년 학사, 2001년 석사학위를 취득하고, 퍼듀대학교에서 산업공학 박사학위를 취득하였다. 2010년부터 이화여자대학교 경영대학에 재직 중이다. 연구분야는 Markov Decision Process, 강화학습, 텍스트분석, 에너지 시스템이다.