완전한 콜드 스타트 문제에서 교차 도메인 추천 시스템

남규현○, 유재성, 채경수
머니브레인
{ngh3053, jaeseongyou, gc}@moneybrain.ai

Cross-Domain Recommendation System in Complete Cold Start Problem

Gyuhyeon Nam○, Jaeseong You, Gyeongsu Chae
MoneyBrain Inc.

요 약

기존의 교차 도메인 추천은 일반적으로 서로 다른 도메인 데이터의 지식 결합이나 지식 공유를 바탕으로 진행된다. 이러한 방식들은 최소 한 개 이상의 도메인 데이터가 필요해서 모든 도메인의 피드백 데이터가 없는 실제 서비스 초기 상황에는 적합하지 않을 수 있다. 따라서 본 논문에서는 서비스 초반 모든 도메인의 피드백 데이터가 없고 콘텐츠 데이터만 존재하는 상황에서 교차 도메인 추천 시스템을 효과적으로 시작하기 위해 텍스트 임베딩, 클러스터링, 프로파일링 및 콘텐츠 기반 필터링을 활용한 추천 시스템 구성을 제안하고자 한다. 평가를 위해 여행지, 지역 축제, 공연을 포함하는 문화 관광 데이터와, 이에 대한 사용자 프로파일링 결과를 바탕으로 추천을 진행하였다. 그 결과, 콘텐츠 임베딩에 대한 유사도를 시각화하여 교차 도메인 아이템 간 유사성을 확인할 수 있었고, 사용자별 추천 결과를 통해 제안한 교차 도메인 추천 시스템이 유의미하게 동작함을 보였다.

주제어: 추천 시스템, 교차 도메인, 콘텐츠 기반 필터링, 클러스터링

1. 서론

추천 시스템에서 도메인은 특정 활동이나 관심분야로 정의할 수 있다. 도메인은 정의에 따라 분류 수준이 달라지는데, 코미디나 스릴러 같은 속성(attribute) 수준이나, 영화나 책같은 유형(type) 수준, 또는 영화나 식당같은 항목(item) 수준에서 도메인을 구분할 수 있다[1]. 교차 도메인 추천 시스템은 이렇게 서로 다른 도메인 사이에서 상품을 추천하는 시스템이다.

일반적으로 교차 도메인 추천 시스템은 지식 결합(knowledge aggregating)이나 지식 공유(knowledge sharing)를 바탕으로 진행된다. 지식 결합 방식은 원천 도메인(source domain)과 대상 도메인(target domain)의 정보를 결합하여 추천하는 방식이고, 지식 공유 방식은 원천 도메인의 피드백 데이터를 바탕으로 대상 도메인을 추천하는 방식이다. 하지만 지식 결합 방식과 지식 공유 방식 모두 하나 이상 도메인의 피드백 데이터가 있어야 적용할 수 있기 때문에 모든 도메인의 데이터가 부족한 상황이면 적용하기 어렵다는 단점이 있다. 실제 서비스 초기 상황에는 모든 도메인의 피드백 데이터가 부족하기 때문에 이러한 기존의 접근 방식은 적합하지 않다.

따라서 본 논문에서는 서비스 초반 모든 도메인의 피드백 데이터가 없고 콘텐츠 데이터만 존재하는 완전한 콜드 스타트(complete cold start) 상황에서 교차 도메인 추천 시스템을 효과적으로 시작하기 위해 텍스트 임베딩, 클러스터링, 프로파일링 및 콘텐츠 기반 필터링을 활용한 추천 시스템 구성을 제안하고자 한다. 텍스트 임베딩과 클러스터링을 사용해 최적의 아이템 후보들을 선택하고 사용자의 취향을 파악하기 위한 프로파일링을 진행한다. 그리고 프로파일링 결과를 바탕으로 콘텐츠 기반 필터링을 이용해 교차 도메인 데이터베이스에서 사용자의 취향과 유사한 아이템들을 추천한다.

실제 시스템 동작을 보이기 위해 문화체육관광부에서 제공하는 문화 관광 데이터를 사용하여 사용자들에게 프로파일링을 진행했고, 추천 결과를 통해 시스템의 동작을 확인하였다.

2. 관련 연구

지식 통합 방법을 사용하는 교차 추천 시스템은 크게 4가지 종류로 나눌 수 있다. (a) 유저 피드백 데이터를 통합하여 단일 도메인처럼 간주하여 추천하는 방법, (b) 추천 시스템을 도메인 개수만큼 설계하고 사용자의 유사도를 비교하여 추천하는 방법, (c) 하나의 추천 시스템으로 교차 도메인을 학습하여 추천하는 방법, (d) 외부적으로 교차 도메인에 대한 지식 네트워크를 구축하여 이를 이용해 추천하는 방법이다[2-5].

지식 공유 방법도 크게 2가지로 나누어 접근할 수 있다. (a) 교차 도메인의 피드백 데이터를 분해하고 공통 잠재 특징을 계산하여 추천하거나, (b) 원천 도메인에서 사용자의 평가 패턴을 검색하여 대상 도메인에 적용하여 추천하는 방법이다[6-8].

지식 통합 방법은 기본적으로 원천 도메인과 대상 도메인의 데이터가 모두 존재한다는 가정하에 시작한다. 그리고 지식 공유 방법은 원천 도메인의 데이터가 존재할 때 사용할 수 있는 방법이다. 두 방법 모두 최소 하나의 도메인 데이터가 필요하기 때문에 교차 도메인 시

스템 초기에 사용하기 적절하지 않다는 한계가 있다.

3. 시스템 구조

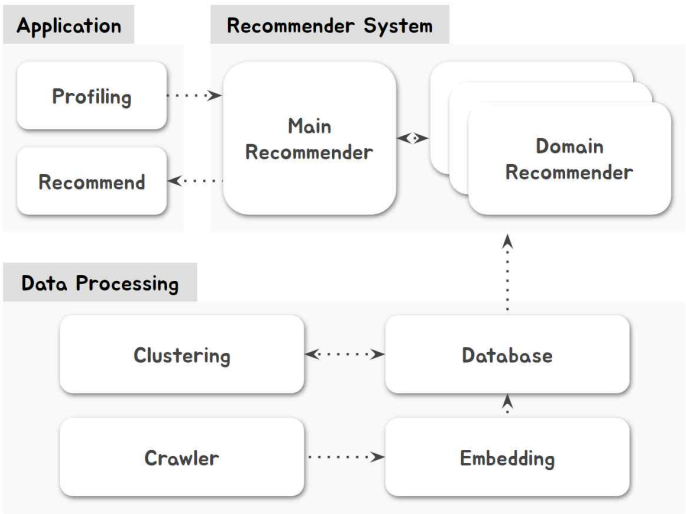

그림 1 교차 도메인 추천 시스템 구조

교차 도메인 추천을 진행하기 위한 전체적인 흐름은 그림1과 같다. 데이터 처리부(그림 하단)는 도메인 데이터를 수집하고 이를 특징 공간으로 임베딩하여 데이터베이스에 저장한 후 클러스터링을 진행한다. 어플리케이션부(그림 상단 좌측)는 사용자의 선호도를 조사하고 추천 결과를 받아와서 사용자에게 제공한다. 마지막으로 추천 시스템부(그림 상단 우측)는 사용자 취향을 바탕으로 콘텐츠 기반 필터링을 적용하여 도메인마다 추천 결과를 산출하고 취합한다. 이번 장의 나머지 부분에서는 시스템 내 각 요소들이 필요한 이유와 구현 방식을 설명한다.

3.1. 텍스트 임베딩

콘텐츠 기반 필터링은 사용자의 피드백 데이터와 아이템에 대한 정보를 바탕으로 사용자가 선호하거나 구입한 아이템과 유사한 아이템을 추천하는 방식이다. 따라서 아이템 간 유사도를 구하기 위해 아이템의 텍스트 데이터를 임베딩하여 유사도를 측정할 필요가 있다.

텍스트 임베딩 방법 중 TF-IDF는 정보 검색 분야에서 널리 사용하는 방법으로, 문서 집합이 있을 때 어떤 단어가 특정 문서 내에서 얼마나 중요한지를 수치로 표현하는 알고리즘이다. 특정 단어가 한 문서에서 자주 등장한다면 중요한 단어일 수 있지만 여러 문서에서 자주 등장한다면 그다지 중요하지 않은 단어가 된다. 이를 반영하여 단어의 가중치를 측정할 필요가 있는데, TF-IDF는 단어 빈도(Term Frequency)와 역문서 빈도(Inverse Document Frequency)를 곱하여 가중치를 계산한다. 단어 빈도는 특정 단어가 한 문서 내에서 등장한 횟수를 계산한 값이고 역문서 빈도는 특정 단어가 모든 문서에서 등장한 횟수의 역수를 계산한 값이다.

TF-IDF를 추천 시스템에 적용하기 위해 모든 도메인의 텍스트 데이터를 문서로 간주하고 TF-IDF 가중치를 학습하였다.

3.2. 클러스터링 및 프로파일링

프로파일링은 아이템에 대한 사용자의 취향 및 선호도를 조사하는 작업이다. 콘텐츠 기반 필터링은 사용자의 과거 아이템 소비 기록을 바탕으로 유사한 아이템을 추천하는 방식이므로, 소비 기록이 없는 경우 프로파일링을 통해 사용자의 선호 아이템을 조사할 필요가 있다. 하지만 선호 아이템을 조사하기 위해서 사용자에게 모든 아이템들을 질의하는 것은 현실적으로 불가능하므로, 아이템 다양성을 최대한 반영하도록 대표 아이템을 선택하여 사용자의 취향을 파악해야 한다.

전문가의 도메인 지식을 바탕으로 아이템들의 분류 체계를 확립하여 대표 아이템들을 선별하는 것이 가장 이상적이지만, 많은 비용과 시간이 소모되기 때문에 자동화 가능한 방법이 필요하다. 따라서 비지도 학습인 클러스터링을 적용하여 아이템을 대표하는 최적의 클러스터를 만드는 작업을 진행하였고, 클러스터의 중심값(centroid)을 선택하여 사용자의 프로파일링을 수행하였다. 실험을 위해 클러스터링 중 대표적인 방법인 K-평균 클러스터링(K-means clustering) 알고리즘을 사용하고 elbow method를 이용해 최적의 K값을 구했다.

3.3. 교차 도메인 추천 시스템

교차 도메인 추천을 위해 추천 시스템을 총괄 추천 시스템과 도메인 추천 시스템으로 구별하였다. 총괄 추천 시스템은 사용자의 선호 아이템을 도메인 추천 시스템들에게 제공하고 각자의 결과를 취합하여 사용자에게 전달하는 역할을 한다. 도메인 추천 시스템은 사용자의 모든 선호 아이템을 입력으로 해당 도메인에 속하는 유사 아이템을 출력하여 총괄 추천 시스템에게 전달한다.

예를 들어, A, B 도메인에 대한 사용자의 선호 아이템을 총괄 추천 시스템이 입력 받으면, 총괄 추천 시스템은 A, B 도메인 선호 아이템 전체를 A도메인 추천 시스템, B도메인 추천 시스템에게 전달한다. A도메인 추천 시스템은 입력 받은 A, B 도메인 선호 아이템을 데이터베이스 안의 A도메인 아이템들과 비교하여 높은 유사도를 가진 아이템을 총괄 추천 시스템에게 전달한다. B도메인 추천 시스템도 마찬가지로 A, B 도메인 선호 아이템을 입력받아 데이터 베이스 안의 B도메인 아이템들과 비교하여 높은 유사도를 가진 아이템을 총괄 추천 시스템에게 전달한다.

아이템간 유사도를 계산할 때 TF-IDF 학습 모델이 텍스트 데이터를 입력 받아 벡터를 출력하여 벡터간 유사도를 계산하도록 하였다. 유사도 측정 방식은 코사인 유사도(Cosine Distance)를 사용하였다.

한편 콘텐츠 기반 필터링은 같은 아이템을 입력하면 항상 같은 결과를 출력한다. 따라서 추천의 다양성을 높이기 위해 총괄 추천 시스템이 사용자 선호 아이템이 속

한 클러스터 내에서 랜덤 샘플링을 거쳐서 도메인 추천 시스템에게 전달하도록 설계하였다.

총괄 추천 시스템은 도메인 추천 시스템에게 결과를 전달받으면 이를 취합하여 어플리케이션부에 전달한다.

4. 실험

이번 장에서는 콜드 스타트 상황에서의 교차 도메인 추천을 위해 교차 도메인을 축제, 여행지, 공연으로 설정하고 실험 결과를 설명한다. 먼저 문화 관광 정보 데이터를 수집하여 텍스트 임베딩 된 아이템들이 실제로 교차 도메인에서 유사성을 지니는지 확인하였다. 그 후, 문화 아이템의 클러스터링 과정을 보이고 사용자 프로파일링을 진행하였다. 마지막으로 단일 도메인에 대한 추천 결과와 교차 도메인에 대한 추천 결과를 순서대로 보였다.

4.1. 문화 도메인 데이터 수집

교차 도메인 콘텐츠를 확보하기 위해서 문화체육관광부[1]에서 제공하는 지역축제, 추천여행지, 문화예술공연정보 데이터를 수집하였다. 수집한 정보는 제목, 설명, 위치, 날짜, 장소, 문의와 기타 메타 데이터이고, 클러스터링과 콘텐츠 기반 추천을 위해 제목과 설명 항목을 사용하였다. 도메인별 아이템 개수는 축제 6814개, 여행지 699개, 공연 20654개이다.

4.2. 텍스트 임베딩 시각화

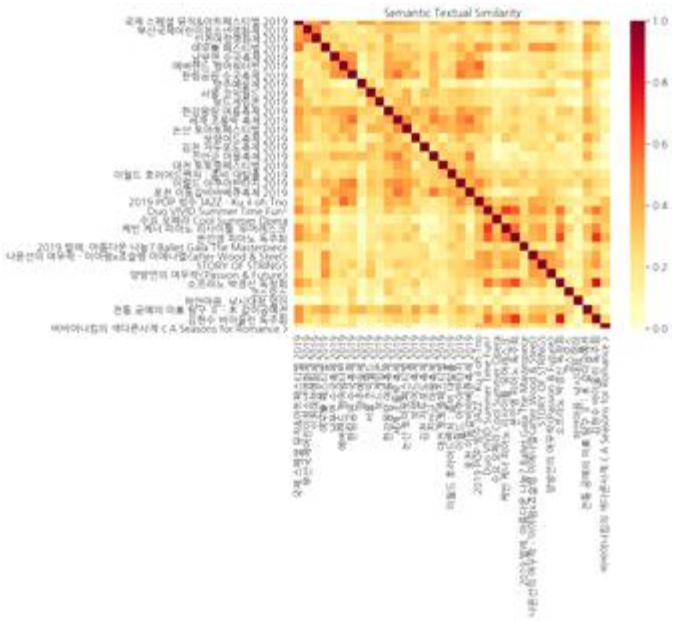

그림 2 교차 도메인 아이템 유사도 시각화

TF-IDF로 임베딩한 문화 콘텐츠들이 실제로 교차 도메인에서 유사성을 가지고 있는지 확인하기 위해 교차 도메인에 속한 문화 아이템의 벡터들을 시각화하였다.

그림2는 축제와 공연 도메인에 대해서 샘플을 추출한 후 유사성을 계산하여 시각화한 결과이다. 값이 0이면 유사도가 낮고, 1이면 유사도가 높은 의미를 가진다. 항목 중에서 제목 마지막에 2019라고 쓰여있는 항목은 축제를, 아닌 항목은 공연을 의미한다. 결과를 보면 대체적으로 축제는 축제끼리, 공연은 공연끼리 연관성이 있는것으로 보인다. 하지만, '2019 POP 빙수 JAZZ Ku il oh Trio' 와 '에버랜드 썸머워터편 2019' 를 보면 서로 다른 도메인이어도 유사도를 가지는 것을 확인할 수 있다.

4.3. 클러스터링 결과 및 사용자 프로파일링

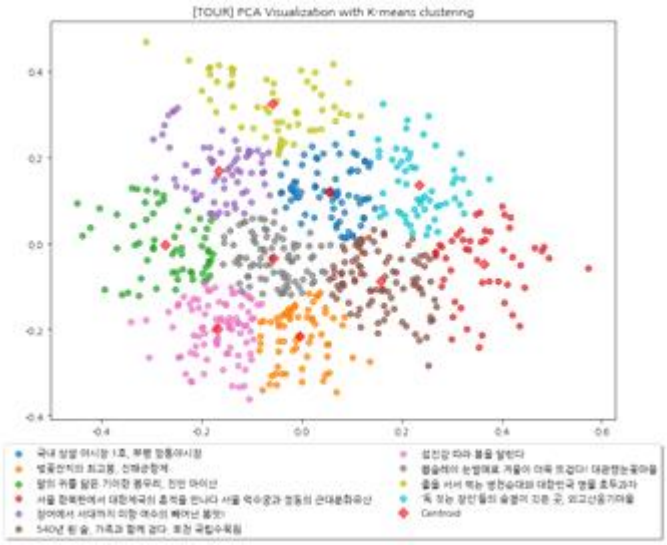

그림 3 여행 도메인 클러스터링 시각화

3가지 문화 도메인에 대해 대표 아이템을 추출하기 위해 클러스터링을 진행했다. 최적의 클러스터링을 위해 각 도메인 마다 최적의 K 값을 계산하였고 평균 10개의 K값이 추정된 것을 확인하였다. 그림3은 여행 도메인에 대한 클러스터링 결과를 나타낸 것이다. 각 클러스터의 중심부분에 해당하는 대표 아이템의 이름을 표기하였고, 총 10 종류의 클러스터를 확인할 수 있다.

클러스터링 결과에 따라 대표 아이템을 산출하여 19명에 대해 선호도를 조사하였다. 조사 방식은 구글 설문지를 이용하여 아이템의 설명 항목을 보여주고, 사용자가 아이템에 대한 선호도를 1점에서 5점까지 선택하는 방식으로 진행하였다. 선호도 점수가 4점 이상이면 그 아이템은 사용자가 선호한다고 가정하고 해당 사용자의 선호 아이템으로 분류하였다.

4.4. 추천 결과

1) https://www.mcst.go.kr/kor/s_culture/festival/festivalList.jsp

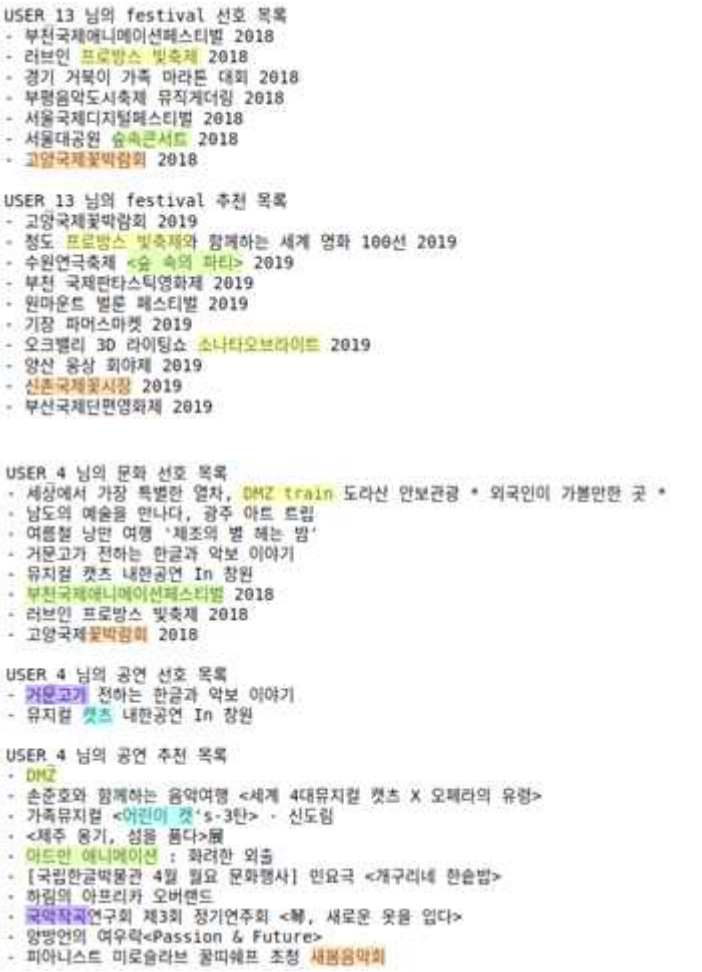

그림 4 사용자별 문화 아이템 추천 결과

그림4는 두 사용자에 대해서 단일 도메인 추천 결과(그림 상단)와 교차 도메인 추천 결과(그림 하단)를 출력한 것이다. 색으로 표시한 아이템은 같은 색끼리 유사도가 높음을 의미 한다.

먼저 단일 도메인 추천 결과를 분석해보면 사용자가 '서울대공원 숲속콘서트'를 선호하면 '수원연극축제 <숲 속의 파티>' 를 추천하고, '러브인 프로방스 빛축제' 를 선호하면 '오크밸리 3D 라이팅쇼 소나타오브라이트' 를 추천하는 것을 볼 수 있다.

다음으로 교차 도메인 추천 결과를 분석해보면, 단일 도메인 결과에서 사용자가 '거문고가 전하는 한글과 악보 이야기'를 선호했다면 '국악작곡연구회 제3회 정기연주회'를 추천하고, '뮤지컬 캣츠' 를 선호하면 '가족뮤지컬 어린이 캣' 을 추천하는 것을 확인할 수 있다. 그 다음 교차 도메인 추천을 확인해보면, 사용자가 축제인 '고양국제꽃박람회'를 선호하면 '피아니스트 미로슬라브 꿀띠쉐프 초청 새봄음악회' 를 추천하고, 여행지인 'DMZ train 도라산 안보관광'을 선호하면 뮤지컬 공연인 'DMZ' 를 추천하는 것을 볼 수 있다.

따라서 텍스트의 유사도를 비교해서 높은 유사도의 아이템을 추천했을 때 단일 도메인과 교차 도메인 모두 의미 있는 결과를 출력하는 것을 확인할 수 있다.

5. 결론

그동안 교차 추천 시스템은 지식 결합이나 공유를 바탕으로 진행 되어 최소 한 개 이상의 도메인의 데이터가 필요했다. 따라서 기존 방식들은 교차 도메인의 데이터가 모두 부족하면 적용이 힘든 단점이 있었다. 교차 추천 시스템을 실제 서비스에 적용하면 모든 도메인에서 콜드 스타트 문제가 발생할 수 있기 때문에 본 논문에서는 클러스터링을 이용한 프로파일링과 콘텐츠 기반 필터링을 적용한 추천 시스템을 제안하였다. 다양한 아이템에 대한 사용자 선호도 프로파일을 확보하기 위해 클러스터 분석을 수행하고, 콘텐츠 기반 필터링을 적용한 시스템으로 교차 도메인의 아이템을 추천했다. 평가를 위해 축제, 여행지, 공연의 문화 데이터를 수집하여 추천을 진행하였고 실제 사용자의 데이터를 토대로 추천 결과를 보였다. 그 결과, TF-IDF 방식의 텍스트 임베딩 방법으로 교차 도메인의 유사성을 파악할 수 있었고, 단일 도메인과 교차 도메인에서 추천 시스템이 유의미하게 동작함을 확인하였다.

감사의 글

이 연구는 기상청 미래유망민간기상서비스성장기술개발(R&D)(1365003036)의 지원으로 수행되었습니다. 또한 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2019년도 문화기술연구개발 지원사업으로 수행되었습니다.

참고문헌

[1] Ivan Cantador, Ignacio Fernandez-Tobıas, Shlomo Berkovsky, and Paolo Cremonesi, "Cross-Domain Recommender Systems", Recommender Systems Handbook, 2015, pp.919-959

[2] Loni, B, Shi, Y, Larson, M. A., Hanjalic, A., "Cross-Domain Collaborative Filtering with Factorization Machines.", Proc. of the 36th European Conf. on Information Retrieval, 2014

[3] Abel, F., Araújo, S., Gao, Q., Houben, G.-J. "Analyzing Cross-system User Modeling on the Social Web.", Proc. of the 11th International Conference on Web Engineering, 2011, pp. 28-43

[4] Berkovsky, S., Kuflik, T., Ricci, F., "Cross-Domain Mediation in Collaborative Filtering", Proc. of the 11th International Conference on User Modeling, 2007, pp.355-359

[5] Givon, S., Lavrenko, V., "Predicting Social-tags for Cold Start Book Recommendations", Proc. of the 3rd ACM Conference on Recommender Systems, 2009, pp.333-336

[6] Pan, W., Xiang, E. W., Yang, Q, "Transfer Learning in Collaborative Filtering with Uncertain Ratings", 2012, pp.662-668

[7] Pan, W., Xiang, E. W., Liu, N. N., Yang, Q., “Transfer Learning in Collaborative Filtering for Sparsity Reduction”, Proc. of the 24th AAAI Conf. on Artificial Intelligence, 2010, pp.210-235
[8] Cremonesi, P., Quadrana, M., “Cross-domain Recommendations without Overlapping Data: Myth or Reality?”, Proc. of the 8th ACM Conference on Recommender Systems, 2014