행렬 분해 기반 협업 필터링을 활용한 콘텐츠 추천 모델에 관한 연구

김경훈, 황동엽, 김기형*
아주대학교
ngngman@ajou.ac.kr, bc8c@naver.com, kkim86@ajou.ac.kr*

A Study on the Contents Recommendation Model Using Matrix factorization Based Collaborative Filtering

Kyung-Hoon Kim, Dong-Yeop Hwang, Ki-Hyung Kim*
Ajou Univ.

요 약

협업 필터링 기반 추천 시스템은 OTT(Over the Top)과 같은 서비스에서 활용되어지고 있다. 그러나 해당 시스템은 새로운 출시된 콘텐츠와 같은 특정 콘텐츠에서 콜드 스타트 문제(Cold-start problems)를 발생시키는 한계점을 갖는다. 이에 본 논문에서는 기존의 협업 필터링 추천 시스템에 행렬 분해 기법을 적용하여 기존의 한계점을 개선한 새로운 콘텐츠 추천 모델을 제안한다. 본 추천 모델은 시청자와 영화를 장르 값을 기준으로 정렬하고 특이값 분해를 이용하여 특징점을 찾아 유사도를 비교해 가장 유사도가 높은 콘텐츠를 시청자에게 추천하는 방식이다. 실험을 통해 기존 협업 필터링 기반 추천시스템의 초반 추천 문제를 해결할 수 있었으며 낮은 컴퓨팅 환경에서도 해당 기법을 적용 할 수 있음을 확인하였다.

Ⅰ. 서 론

신종 코로나바이러스 감염증(코로나19)의 세계적 대유행이 과속화되면서 '언택트(비대면)' 산업이 부상하고 생활 속에 광범위하게 확산되고 있다. 그 중 시간, 장소의 제약 없이 원하는 콘텐츠를 제공하는 OTT(Over the Top) 서비스는 올해 1분기에 가입자 수가 1억 8000만명을 넘는 실적을 보이며 끝없는 인기를 얻고 있다. OTT 서비스의 중심에는 일정한 규칙에 따라 사용자가 좋아할 만한 콘텐츠를 추천해주는 알고리즘이 자리잡고 있으며, 이는 콘텐츠에 대한 장르, 카테고리, 평점 등을 이용하여 콘텐츠 기반 또는 아이템 기반으로 추천을 수행한다.

그러나 일정 투표수 이하로 발생하는 평점 정보의 불확실성으로 인한 콜드 스타트 문제(Cold-start problems)[1], 특정 사용자의 취향을 고려하지 못하고 유명한 상품이 자주 추천되는 문제[2] 등이 발생한다. 이에 본 논문에서는 행렬 분해 기반 협업필터링을 활용하여 사용자간 콘텐츠 소비 취향을 분석하고 비슷한 패턴을 가진 사용자에게 소비하고자 하는 콘텐츠 추천 모델을 제시한다.

논문의 구성은 다음과 같다. 2장에서는 협업 필터링 추천 시스템과 IMDb 사용자등급, 행렬 분해에 대하여 설명한다. 다음으로 3장에서는 본 논문에서 제안하는 "행렬 분해 기반 협업 필터링을 활용한 콘텐츠 추천 모델"에 대해 설명한다. 4장에서는 실험을 통하여 해당 시스템에 대하여 분석한다. 마지막으로 5장에서는 본 논문에 대한 결론을 제시한다.

Ⅱ. 관련연구

1) 협업 필터링 추천 시스템(Collaborative Filtering Recommendation System)

협업 필터링 추천 시스템은 자신과 취향이 비슷한 사람을 하나로 묶어 같은 그룹에서 다른 사람이 시청한 영화가 나에게도 흥미가 있을 것이라는 가정에서 출발한다. 그러나 그룹 내 평가 내역을 이용하기 위해서는 충분한 양의 데이터가 수집되어야하기 때문에 시스템 초반에는 정확한 추천이 어렵다.[3]

2) IMDb 사용자 등급(Internet Movie Database User Rating)

IMDb는 세계에서 가장 큰 영화 데이터베이스이다. IMDb는 다양한 필터를 적용하고 특정 가중치를 부여하여 독자적인 등급 계산법으로 영화에 대한 등급을 제시한다. 계산된 등급은 WR(Weighted Rating)이며 계산식은 다음과 같다.[4]

$$WR = (\frac{v}{v+m} \cdot R) + (\frac{v}{v+m} \cdot C)$$

- v : 영화의 득표 수
- m : 특정 순위안의 최소 투표 수
- R : 영화의 평균 평점
- C : 전체 차트의 평균 평점

3) 행렬 분해(Matrix factorization)

m명의 사용자, n개의 콘텐츠가 주어졌을 때 사용자-콘텐츠 평점 행렬 $A \in R^{m \times n}$는 다음과 같이 정의된다. 그러나 실제로 모든 사용자가 모든 콘텐츠에 평점을 매기지 않았기 때문에 사용자-콘텐츠 평점 행렬 A는 비어있는 원소가 다수 존재할 것이다. 이를 해결하고자 특정 계수를 통해 불완전한 사용자-콘텐츠 평점 행렬 A의 비어있는 원소 값을 추측하는 것이 행렬 분해법의 목표이다.[5]

머신러닝에서 행렬 분해가 선호되는 이유는 모든 잠재 요인이 양수로 표현되기 때문에 해석이 쉽고, 기존 특이값 분해에 비해 훨씬 희소한 행렬을 만들어내는 경향이 있기 때문이다. 희소한 행렬은 메모리를 덜 차지하면서 서비스 시에 계산할 양도 줄어들기 때문에 실용적으로 큰 장점이라고 할 수 있다.[6]

Ⅲ. 제안 방법

본 논문에서는 기존의 협업 필터링 추천 시스템의 문제점을 개선하기 위해 행렬 분해 기법을 적용시켜 콜드 스타트 문제를 해결하는 추천 시스템을 제안한다. 제안 방법은 사용자와 콘텐츠간의 행렬 데이터와 잠재 요인

을 결합하고 두 개의 특성 행렬을 이용하여 결측치(Missing Values)를 예측한다. 여기서 잠재 요인을 영화 콘텐츠 데이터의 장르 값으로 정의하였다.

먼저 개인 투표 기록과 영화 콘텐츠 데이터를 장르 값을 기준으로 '사용자-장르', '영화-장르'로 테이블을 생성한다.

이 후 행렬 분해의 방식 중 $m \times n$ 행렬의 고유값과 고유벡터를 이용하여 행렬을 특정한 구조로 분해하여 신호처리와 통계학 등의 분야에서 자주 사용되는 특이값 분해(Singular Value Decomposition; SVD)를 이용하여 추출된 영화의 특징 값을 추출한다.

마지막으로 유사도를 비교하여 그 중 유사도가 높은 콘텐츠를 선정해 시청자에게 추천한다. 이 때 두 벡터간의 각도의 코사인 값을 이용하여 유사도를 구하는 코사인 유사도(Cosine similarity)를 사용하여 벡터 간 유사도를 계산하였다.[7]

Ⅳ. 실험 및 결과

1) 실험 데이터

실험 데이터는 빅데이터 솔루션 대회 플랫폼인 'kaggle'의 The Movies Dataset의 데이터로 130,000명의 개인 투표 기록(12,528,190개)과 45,466개의 영화 콘텐츠 데이터를 사용하였다. 영화 콘텐츠 데이터 중 평점 값(ratings)은 IMDb(Internet Movie Database)의 영화 평가 방식인 WR(Weighted Rating)을 이용하여 평점 값을 재 정의하였다. 46,482개의 영화 데이터 중 약 10%의 순위인 5,000위까지 최소 투표수를 기준으로 WR을 계산하였으며, 이는 전체 영화 투표수의 약 10%에 해당된다.

2) 실험

먼저 '사용자-장르', '영화-장르'로 테이블을 생성하고 모든 값을 0으로 채워 행렬 초기화 작업을 실시한다.

이 후 Scikit-learn에서 제공하는 특이값 분해(SVD)를 이용하여 4,957개의 영화에서 12개의 특이 값을 추출한다. 마지막으로 코사인 유사도를 사용하여 유사도를 측정하여 모델을 생성한다.

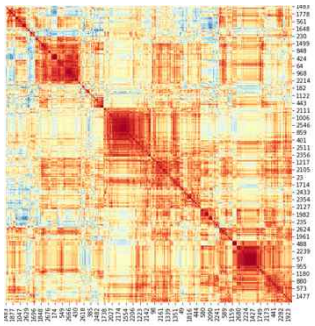

그림 1. 콘텐츠 간 유사도 판단

3) 결과

그림 2는 'Secret Window' 영화의 유사도가 97% 이상인 콘텐츠의 행렬 분해 기반 협업 필터링 결과 값이다. 데이터의 양이 증가할수록 많은 콘텐츠를 추천해주는 협업 필터링 추천 시스템의 기본적 특성은 존재하나 적은 데이터에서도 높은 유사도의 콘텐츠를 추출하여 협업 필터링 추천 시스템의 초반 추천 문제를 해결 할 수 있었다.

그림 2. 행렬 분해 기반 협업 필터링 결과

또한, 행렬 분해 기법을 사용함으로써 기존의 협업 필터링 추천 시스템의 행렬 과다 생성 문제를 해결 할 수 있었다. 이는 컴퓨터 성능 저하 문제를 해결할 수 있어 낮은 컴퓨팅 환경에서도 많은 결과 값을 추출할 수 있을 것으로 예상된다.

Ⅴ. 결론

본 논문에서는 행렬 분해 기반 협업 필터링을 활용한 콘텐츠 추천 모델을 제안하였다. 이는 영화 데이터의 장르 값을 기준으로 시청자와 영화간의 특이점을 분류하여 적은 투표 기록에서도 시청자에게 영화를 추천해주는 효과를 갖는다. 그러나 협업 필터링 추천 시스템의 기본적인 한계를 포함하고 있기 때문에 향후 연구에 있어 행렬 분해 기반 협업 필터링 모델과 콘텐츠 기반 추천 모델을 결합한다면 더욱 정확한 예측을 할 수 있을 것으로 판단된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업과 2018년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업의 지원과 2020년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (IITP-2020-2018-0-01396, NRF-2018R1D1A1B07048697, P0008703, 2020년 산업전문인력역량강화사업)

참 고 문 헌

[1] Hyung Jun Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," Information Science, pp. 67-51, 2008.

[2] Yoon-Joo Park, Alexander Tuzhilin, "The Long Tail of Recommender Systems and Hot to Leverage It," Recommender System, pp. 11-18, 2008.

[3] 유상원, 이홍래, 이형동, 김형주, "TV프로그램을 위한 내용기반 추천 시스템," 정보과학회논문지, pp. 638-692, 2003.

[4] Ping-Yu Hsu, Yuan-Hong Shen, and Xiang-An Xie, "Predicting Movies User Ratings with Imdb Attributes," Rough Sets and Knowledge Technology, pp. 444-453, 2014.

[5] 손동희, 심규석, "문자 수준 컨볼루션 뉴럴 네트워크를 이용한 추천시스템에서의 행렬 분해법 개선," 정보과학회 컴퓨팅의 실제 논문지, pp. 93-98, 2018

[6] 유찬우, 김희천, "개인화 추천 시스템의 머신러닝을 위한 음수 미포함 행렬 분해와 변수 선택 기법의 비교," 디지털콘텐츠학회논문지, pp. 793-798, 2020

[7] 방한별, 이혜우, 이지형, "연령 및 프로그램 줄거리를 활용한 콘텐츠 기반 TV 프로그램 추천 시스템," 한국컴퓨터정보학회, pp. 51-54, 2015