한국빅데이터학회지
제7권 제1호, 2022, pp. 187-192 https://doi.org/10.36498/kbigdt.2022.7.1.187

협업 필터링을 활용한 비교과 프로그램 추천 기법: C대학 적용사례

Non-Curriculum Recommendation Techniques Using Collaborative Filtering for C University

전유정[1] · 양경은[2] · 조완섭[2†]

충북대학교 빅데이터협동과정[1], 충북대학교 경영정보학과[2]

요 약

많은 대학교에서 다양한 교과 및 비교과 활동을 통해 학생들의 취업 역량을 향상하기 위해 노력하고 있지만, 취업을 준비하는 학생마다 목표와 하고자 하는 활동이 다르다. 따라서 기존에 획일적이고 종합적으로 제공하고 있는 프로그램이 실제로 학생들에게 적합한지 여부를 판단하기 어려우므로 개인화 추천 시스템의 도입이 필요하다. 본 연구에서는 충북대학교의 모든 학생에게 일괄적으로 제안되고 있는 비교과 프로그램을 학년 및 학과별로 분류하여 제시하는 방법을 제안하였다. 또한, 비교과 프로그램에 참여한 학생의 평점 데이터를 사용하여 협업 필터링 모델 3가지를 구현하고, 성능을 비교해 가장 정확도가 높은 모델로 개인화된 맞춤형 추천을 제안한다.

■ 중심어 : 머신러닝, 추천 시스템, 분류, 협업 필터링, KcBERT

Abstract

Many schools are trying to improve students' competencies through many subjects and non-curricular activities, each students has different goals and different activities to prepare for employment. Accordingly, it is difficult to determine whether the programs offered in a comprehensive and comprehensive manner in the existing subject and non-curricular subjects systems are actually suitable for students, so it is necessary to introduce a personalized system. In this study, a method was proposed to classify non-departmental subjects that are uniformly provided to all students of Chungbuk National University by grade level and department. In addition, three types of collaborative filtering models are implemented using the evaluation score of students who participated in the non-curricular program, and personalized recommendations are proposed with the most accurate model by comparing performance.

■ Keyword : Machine learning, Recommendation system, Classification, Collaborative filtering, KcBERT

2022년 05월 13일 접수; 2022년 05월 24일 게재 확정.
* 본 연구는 참고문헌[6]을 토대로 작성되었음.
† 교신저자 (wscho63@gmail.com)

Ⅰ. 연구의 필요성

수많은 비교과 프로그램들 가운데 학생 개인에게 적절한 프로그램을 선택하기 위해서 콘텐츠의 내용과 참여 대상에 대한 고려가 필요하다. 이를 위해 학과, 학년, 프로그램 평점 등과 같은 정보를 참고해 해당 학생이 참여할 것으로 예상되는 프로그램을 제시한다면 학생의 관심과 참여도를 높일 수 있을 것으로 판단된다.

본 연구는 학생들이 재학 기간 취업 준비 및 역량 향상을 위해 비교과 프로그램을 적극적으로 활용할 수 있도록 기존 시스템에 맞춤형 추천 도입을 제안한다. 충북의 C 대학교에서 지원하는 비교과 프로그램 관리 사이트 '씨앗'에서 수집된 데이터를 바탕으로 협업 필터링(Collaborative Filtering, CF) 기법을 적용해보고, 협업 필터링을 사용할 수 없는 새로운 프로그램의 경우 프로그램 제목에 따른 참여 대상의 학과 및 학년 분류(Classification) 모델을 적용해 초기 데이터가 없는 문제를 보완하고자 한다.

협업 필터링 기법 중 유사도 기반 필터링(Memory-based Collaborative Filtering)은 아이템 기반 협업 필터링(Item-based Collaborative Filtering), 사용자 기반 협업 필터링(User-based Collaborative Filtering)의 두 가지 기법으로 구성되며, 잠재 요인 기반의 협업 필터링(Latent Factor Collaborative Filtering)은 대표적으로 행렬 인수 분해(Matrix Factorization, MF)기법이 있다. 이 세 가지 필터링 알고리즘을 통해 각 모델의 정확도를 비교하여 '씨앗' 사용자 데이터에 가장 적절한 기법을 채택한다.

Ⅱ. 관련 연구

협업 필터링이란 많은 사용자로부터 수집한 취향 정보를 바탕으로 사용자들의 관심사가 될 항목을 예측해주는 기법이다.

김두형 등(2020)은 협업 필터링을 활용한 교과목 추천 시스템을 구축하였다. 재학생과 졸업생 수강 이력 데이터와 강의평가 점수를 바탕으로 코사인 유사도를 측정하고 근접한 이웃을 찾아 교양과목을 추천하는 서비스를 구현했다.

최재용 등(2021)의 연구에서 학생이 희망하는 진로에 적절한 교과 및 비교과 활동 선택에 도움 주고자 학생의 교내 활동들에 대한 평점, 댓글, 콘텐츠 검색 등의 온라인 활동 정보를 수집했다. 도출된 특정 키워드나 콘텐츠와 관련된 기업들을 학생의 관심군에 속한 기업으로 간주하였다. 이를 바탕으로 졸업생들의 신뢰도와 유사도를 측정해 참고 모델이 되는 졸업생을 선정하고, 졸업생 데이터와 유사한 교과, 비교과 등의 활동을 협업 필터링을 통해 추천하였다.

기존 연구들은 졸업생 데이터가 존재하기 때문에 이를 바탕으로 협업 필터링을 수행해 교과 및 비교과 콘텐츠를 추천한다.

하지만 본 논문은 비교과목 추천에만 초점을 두고, 재학생 참여 이력이 전무한 새로운 프로그램을 학생에게 제안하는 방법과 졸업생데이터와 같이 참고할 데이터가 없는 상황에서 재학생 참여 이력만을 바탕으로 추천시스템을 적용하여 연구한다.

Ⅲ. 연구 방법

3.1 학년 및 학과 분류를 위한 데이터

본 연구의 실험에 사용된 데이터는 충북대학교 씨앗 홈페이지가 개설된 2020년 3월부터 2021년 12월 초까지 등록된 비교과 활동 목록 중 프로그램 제목과 역량 항목, 그리고 모집 기간, 참여 대상의 학년 및 학과(단과대학)를 크롤링해 프로그램 '제목'과 분류 대상이 되는 '학년', '학과' 열로 구성된 1,345행 데이터이다.

3.2 맞춤형 추천을 위한 사용자 데이터

협업 필터링 실험을 위해 '씨앗' 사용자 100명의 활동 내용 데이터 1,222개를 수집하였다. <그림 1>은 학생들이 참여한 프로그램 데이터이며 제목인 'title', 프로그램 제목 번호 'programId', 비식별화한 사용자 고유번호 'userId', 학과(단과대학) 'department', 고학년, 저학년, 전체 학년을 구분하는 'level' 그리고 프로그램 참여 후 사용자가 부여한 평가점수인 'rating'의 6열로 구성되어 있다.

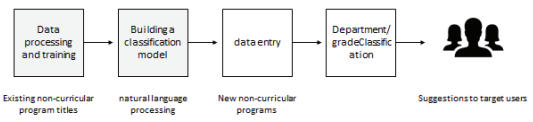
	userId	departaent	level	title	programId	rating
0	1	경영대학	저학년	취업특화 경영대 학생 진출가능 직무탐색 및 분석특강	13	4
1	2	경영대학	저학년	취업특화 경영대 학생 진출가능 직무탐색 및 분석특강	13	5
2	4	경영대학	고학년	취업특화 경영대 학생 진출가능 직무탐색 및 분석특강	13	3
3	14	사범대학	저학년	취업특화 경영대 학생 진출가능 직무탐색 및 분석특강	13	2
4	18	생활과학	저학년	취업특화 경영대 학생 진출가능 직무탐색 및 분석특강	13	2
...	...	...	...	...	...	...
1217	100	전자정보대학	고학년	전공 및 졸업논문 학술발표대회 1차	788	5
1218	100	전자정보대학	고학년	2021학년도 대학혁신지원사업 비교과 프로그램 CK ICT 4차 산업혁명 특강 소...	175	4
1219	100	전자정보대학	고학년	소프트웨어학과 미래설계준비 학습법 특강 II	221	3
1220	100	전자정보대학	고학년	2021년 2학기 창업특강 인공지능을 만드는 공장이 만들 미래	222	4
1221	100	전자정보대학	고학년	소프트웨어학과 창업설계 창업특강 2회차	238	5

1222 rows × 6 columns

〈그림 1〉 '씨앗' 사용자의 프로그램 참여 데이터

3.3 비교과 프로그램 추천 모델 구성

비교과 추천 모델의 구축을 위해 두 가지 모델을 제안한다. 첫 번째는 프로그램의 참여 대상을 NLP 기반의 응용 모델로 학습된 데이터를 통해 분류하는 모델이다. 기존에 진행되었던 프로그램 제목을 바탕으로 제목의 명사 키워드를 추출하고 토큰화 후 특정 키워드가 어떤 학생 집단에 추천하기 적절한 활동인지 분류하는 학습을 진행한다. 이는 추천 시스템을 도입해 운영할 시 새로운 프로그램이 입력값으로 들어오게 되었을 때, 해당 프로그램에 대한 참여 이력이 전혀 없는 콜드 스타트 문제의 상황에도 적절한 참여 대상을 분류할 수 있는 모델이 된다. <그림 2>는 분류 모델의 과정을 표현한 것이다.

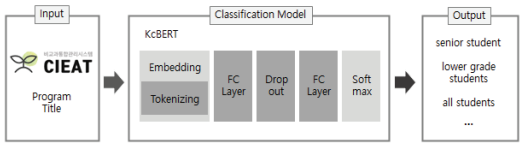


〈그림 2〉 분류 모델 절차

두 번째는 협업 필터링 기법을 적용하여 추천하는 모델이다. 이 모델은 프로그램에 참여한 학생들의 데이터를 바탕으로 코사인 유사도로 계산한 최근접 이웃 기반 필터링과 잠재 요인 분석의 행렬 분해 기반의 두 가지 방법으로 실험하고 추천 결과를 평균 제곱근 오차(Root Mean Square Error, RMSE) 값으로 비교하여 더 좋은 성능의 모델을 채택한다.

3.4 비교과 프로그램별 분류 모델 구현

비교과 프로그램에 대하여 학생들의 참여 명세가 전혀 없거나 부족한 경우를 위한 모델이다. 비교과 프로그램명에 따라 참여 대상을 어느 학과(단과대학) 및 학년인지 예측할 수 있도록 모델을 구현한다. 입력값은 비교과 프로그램의 제목이며 활동의 대상을 예측할 훈련 데이터이다. 실험에는 비교과 프로그램을 지원하는 '씨앗' 사이트에서 수집한 1,345개의 크롤링 데이터를 사용한다. <그림 3>은 분류 모델을 위한 BERT 학습 모델의 절차이다.



〈그림 3〉 BERT Learning Model

분석 언어는 python을 사용하였고 KcBERT를 통해 프로그램 제목을 형태소 분리 후 명사 위주로 학습시켰다. KcBERT는 네이버 댓글을 통해 학습하여 신조어를 포함하는 BERT 모델이다(이준범, 2020). 실험에 사용된 프로그램 제목 데이터 일부에서 신조어 및 줄임말을 사용하는 것이 확인되었기에 이 모델로 채택하였다. 또한

Pytorch Lightning 라이브러리를 이용해 저학년, 고학년, 전체 학년으로 분류하는 학년 예측 모델과 단과 대학별로 분류하는 학과 예측의 2가지 모델로 구축하였다. 분류 라벨 수에는 학년 클래스 3개, 학과 클래스 12개를 각 모델에 입력한다. 입력 데이터는 KcbertTokenizer을 통해 벡터값으로 출력된다. 제목 데이터에서 명사 토큰으로 만들어진 벡터 데이터는 차례대로 KcBERT 모델의 3개 계층을 통해 훈련된다.

3.5 비교과 프로그램 추천 모델 구현

학생들이 비교과 프로그램에 참여한 후 부여한 평점과 같은 선호도를 바탕으로 만들어진 취향 정보 데이터를 통해 다른 비교과 프로그램을 추천하는 모델이다. 다양한 학생들의 비교과 프로그램 참여 데이터를 사용하여 협업 필터링 기법을 적용하고 참여하지 않은 프로그램들에 대한 선호도를 예측해 높은 유사도를 보이는 프로그램을 추천한다. 협업 필터링의 대표적인 세 가지 기법을 모두 적용해 본 후 결과 비교를 통해 가장 성능이 좋은 모델을 채택한다.

3.5.1 아이템 기반 비교과 프로그램 추천 모델

아이템 기반 추천 모델은 학생들이 비교과 프로그램에 참여한 후 평가한 점수를 바탕으로 프로그램 간의 유사도를 계산하고, 특정 프로그램을 수강한 학생이 만족도가 높을 때, 그와 유사한 만족도로 예측되는 다른 프로그램을 추천하는 모델이다. 사용자 간 유사도는 코사인 유사도를 사용하여 이웃들을 찾고, Top-N 유사도를 가지는 벡터만을 가지고 해당 학생에게 추천하도록 학습한다. Top-N 추천이란 사용자의 선호가 예상되는 상위 N개를 추천한다는 개념이다(이희춘, 2006).

3.5.2 사용자 기반 비교과 프로그램 추천 모델

사용자 기반 추천 모델은 학생들 사이의 프로그램 선호도를 분석하여 해당 학생과 유사한 성향의 학생들이 좋은 평점을 준 프로그램을 추천하는 모델이다. 마찬가지로 코사인 유사도 연산을 통하여 모든 사람의 유사도를 검색하고, Top-N 유사도를 가지는 학생에 대한 프로그램의 예상 평가점수 산출한다.

3.5.3 잠재 요인 기반 비교과 프로그램 추천 모델

잠재 요인 분석 기법은 행렬 분해 방식 중 특잇값 분해를 이용하여 성능이 우수하다고 알려져 있다. 실험에는 surprise 라이브러리를 사용한다. 실험은 데이터 로딩, 모델 설정 및 학습, 예측 및 평가의 단계로 진행한다.

세 모델의 모든 학습이 끝나면 결과를 RMSE 값으로 성능을 확인한다.

Ⅳ. 성능평가

4.1 비교과 프로그램 분류 모델의 훈련 결과

1,345개의 데이터 중 80%는 학습 데이터로 사용했으며, 20%는 테스트 데이터로 사용하였다. 평가 지표는 정밀도(precision), 재현율(recall), F1 score, 정확도(accuracy)를 사용하였고, 지지도(support)도 확인하였다.

학과 모델 검증 세트의 훈련 결과 정확도는 0.8848로 비교적 높은 정확도를 보였으며, val_f1은 0.6730, val_loss는 0.0138, val_precision은 0.7349, 그리고 val_recall은 0.6790의 결과를 확인하였다.

고학년, 저학년, 전체 학년으로 나눈 학년 모델 검증 세트의 훈련 결과는 다음과 같다. 훈련의 정확도는 0.8921의 비교적 높은 정확도를 보였으며, val_f1은 0.8598, val_loss는 0.0322, val_precision은 0.8381, 그리고 val_recall은 0.8865의 결과가 나왔다.

4.2 협업 필터링 기반 추천 결과

아이템 기반 필터링을 통한 모델은 평점 예측을 통해 가장 높은 점수를 부여할 것 같은 프로그램 상위 10가지를 도출하고, 학생이 이미 참가하고 평점을 매긴 프로그램을 제외한 리스트를 다시 반환해 유사도가 높은 상위 10가지 프로그램을 도출했다. 40번 학생을 기준으로 확인한 결과는 <그림 4>와 같다.

title	pred_score
취업패키지 923	1.567224
스트레스 프리 마음챙김 명상 집단상담 프로그램	1.170081
수요조사 2021 입사지원서 작성 캠프	1.025039
사회과학대학 단기 특강 PPT 엑셀 영상편집 R코딩	0.919928
비즈니스 엑셀2 3 급 자격 과정	0.850004
2020 엑셀데이터분석 2 3 급 자격 과정	0.798625
코로나 시대의 사회적 약자와 진보 정치	0.635902
엑셀함수&단축키 활용법	0.609985
1차 ZOOM 온라인 학습법 특강	0.532822
2021년도 대학혁신지원사업 비교과 프로그램 CK ICT 4차 산업혁명 특강 소프트웨어학과	0.517286

〈그림 4〉 참여 프로그램을 제외한 필터링 결과

협업 필터링 세 모델의 평점 예측 결과를 비교한 결과, TOP-20를 적용한 모델의 성능이 조금 더 높은 정확도를 보이며, 유사도 기반의 아이템 및 사용자 필터링 기법과 행렬 분해 기반의 SVD 기법의 정확도를 비교한 결과, 아이템 기반 인접 TOP-20 이웃 RMSE 값은 2.1031, 사용자 기반 인접 TOP-20 이웃 RMSE 값은 2.3880 그리고 SVD 기법은 0.7069로 잠재 요인 기법의 SVD의 정확도가 가장 높음을 확인했다.

Ⅴ. 결론 및 제언

본 논문에서는 충북의 C 대학교에서 지원하는 비교과 프로그램 서비스에 맞춤형 추천 기법의 도입을 제안하고자 실험을 진행했다. 비교과 프로그램의 제목과 목표 대상을 분류한 결과 학과(단과대학) 예측 모델은 88%, 학년 예측 모델은 89%의 정확도를 보여 프로그램 제목에 따른 학과 및 학년 분류의 가능성을 확인했다. 또한, 협업 필터링의 최근접 이웃 기반의 유사도 분석과 잠재 요인 기반의 특잇값 분해 결과, 아이템 기반 인접 TOP-20 이웃 모델의 RMSE 값은 2.1031, 사용자 기반 인접 TOP-20 이웃 모델의 RMSE 값은 2.3880 그리고 SVD 기법 모델의 RMSE 값은 0.7069로 '씨앗' 사이트 내에서 비교과 프로그램 추천 시 잠재 요인 협업 필터링 모델을 채택하는 것이 적절하고 할 수 있다.

한계점은 다음 같다. 첫째, 비교과 프로그램을 추천하는 데 있어 학과(단과대학)와 학년으로 분류하는 기준이 기존에 진행되었던 비교과 프로그램명이라는 점이다. 새로운 인풋 제목 데이터가 들어왔을 때, 기존 프로그램명에는 존재하지 않아 학습하지 못한 키워드가 있다면 분류 정확도가 떨어진다는 문제점이 있다. 이를 극복하기 위해서 비교과 프로그램별 핵심 키워드와 내용을 명시하여 정확도를 높이는 방안이 필요하며, 이를 활용한다면 콘텐츠 기반 필터링에 의한 분류 및 추천을 진행할 수 있다.

둘째, 분류 모델에 사용한 데이터의 불균형 문제가 있다. 학과별로 다른 데이터 크기 때문에 프로그램 수가 적은 경영대학과 사회과학대학에 대한 훈련이 제대로 이루어지지 않음을 알 수 있다. 이를 보완하기 위해 모델 훈련 전 데이터 크기의 균형을 맞추는 사전 처리가 필요하다.

향후 이 두 모델을 바탕으로 진행한 추천에 대한 사용자의 반응을 확인하고 이를 결과 데이터로 수집하여 다른 필터링 기법과 결합한 하이브리드 기법으로 추천을 고도화하는 연구를 진행하고자 하며, 한계를 보완해 모델을 개선하고자 한다.

참 고 문 헌

[1] 김두형, 신우석, 한기웅, 이진숙, 문기범, 이수강, 한수연, 권혜정, 한성원. “협업필터링을 활용한 대학 교양과목 추천 시스템”, 대한산업공학회 추계학술대회 논문집, 2551-2556, 2020.

[2] 최재용, 임종태, 오영호, 편도웅, 이소민, 백연희, 신보경, 박수빈, 복경수, 유재수. “개인 맞춤형 대학 교육 콘텐츠 추천 시스템의 설계 및 구현”, 한국정보과학회, 168-170, 2020.

[3] 이준범. “KcBERT: 한국어 댓글로 학습한 BERT”, 한글 및 한국어 정보처리 학술대회 논문집, 32, 437-440, 2020.

[4] 이희춘. “추천시스템에서 Top-N 추천을 위한 순위 적합에 관한 연구”, Journal of The Korean Data Analysis Society. 8(6). 2597-2607, 2006.

[5] 전유정, 협업필터링을 통한 비교과 프로그램 추천 시스템, C대학을 중심으로, 석사학위논문, 충북대학교, 2022.

저 자 소 개

전 유 정 (yujung Janu)

·2018년: 청주대학교 광고홍보학 (학사)

·2022년: 충북대학교 빅데이터협동과정 (석사)

·관심분야: 빅데이터, 머신러닝, 추천 시스템 등

양 경 은 (Kyungeun Yang)

·2010년: 충북대학교 경영정보학과 (석사)

·2013년: 충북대학교 경영정보학과 박사과정 (수료)

·관심분야: 빅데이터, 비즈니스 인텔리전스, 데이터 분석 등

조 완 섭 (Wan-Sup Cho)

·1987년: KAIST 전산학과 (박사)

·1996년 ~ 현재: 충북대학교 교수

·관심분야: 빅데이터, 블록체인, 빅데이터거버넌스