ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 September: 27(3): 95~112 http://dx.doi.org/10.13088/jiis.2021.27.3.095

네트워크 분석을 활용한 딥러닝 기반 전공과목 추천 시스템

이재규
연세대학교 산업공학과
(dlworb1994@yonsei.ac.kr)

박희성
연세대학교 산업공학과
(gregmark@naver.com)

김우주
연세대학교 산업공학과
(wkim@yonsei.ac.kr)

대학 교육에 있어서 전공과목의 선택은 학생들의 진로에 중요한 역할을 한다. 하지만, 산업의 변화에 발맞춰 대학 교육도 학과별 전공과목의 분야가 다양해지고 그 수가 많아지고 있다. 이에 학생들은 본인의 진로에 맞게 수업을 선택하여 수강하는 것에 어려움을 겪고 있다. 본 연구는 대학 전공과목 추천 모델을 제시함으로써 개인 맞춤형 교육을 실현하고 학생들의 교육만족도를 제고하고자 한다. 모델 연구에는 대학교 학부생들의 2015년~2017년 수강 이력 데이터를 활용하였으며, 메타데이터로는 학생과 수업의 전공 명을 사용했다. 수강 이력 데이터는 컨텐츠 소비 여부만을 나타낸 암시적 피드백 데이터로, 수업에 대한 선호도를 반영한 것이 아니다. 따라서 학생과 수업의 특성을 나타내는 임베딩 벡터를 도출했을 시, 표현력이 낮다. 본 연구는 이러한 문제점에 착안하여, 네트워크 분석을 통해 학생, 수업의 벡터를 생성하고 이를 모델의 입력 값으로 활용하는 Net-NeuMF 모델을 제시한다. 모델은 암시적 피드백을 가진 데이터를 이용한 대표적인 모델인 원-핫 벡터를 이용하는 NeuMF의 구조를 기반으로 하였다. 모델의 입력 벡터는 네트워크 분석을 통해 학생과 수업의 특성을 나타낼 수 있도록 생성하였다. 학생을 표현하는 벡터를 생성하기 위해, 각 학생을 노드로 설정하고 엣지는 두 학생이 같은 수업을 수강한 경우 가중치를 가지고 연결되도록 설계했다. 마찬가지로 수업을 표현하는 벡터를 생성하기 위해 각 수업을 노드로 설정하고 엣지는 공통으로 수강한 학생이 있는 경우 연결시켰다. 이에 각 노드의 특성을 수치화 하는 표현 학습 방법론인 Node2Vec을 이용하였다. 모델의 평가를 위해 추천 시스템에서 주로 활용하는 지표 4가지를 사용하였고, 임베딩 차원이 모델에 미치는 영향을 분석하기 위해 3가지 다른 차원에 대한 실험을 진행하였다. 그 결과 기존 NeuMF 구조에서 원-핫 벡터를 이용하였을 때보다 차원과 관계없이 평가지표에서 좋은 성능을 보였다. 이에 본 연구는 학생(사용자)와 수업(아이템)의 네트워크를 이용해 기존 원-핫 임베딩 보다 표현력을 높였다는 점, 모델을 구성하는 각 구조의 특성에 맞도록 임베딩 벡터를 활용하였다는 점, 그리고 기존의 방법론에 비해 다양한 종류의 평가지표에서 좋은 성능을 보였다는 점을 기여점으로 가지고 있다.

주제어 : 교육 빅데이터, 노드 임베딩, 네트워크 분석, 딥러닝, 추천 시스템

논문접수일 : 2021년 5월 22일 논문수정일 : 2021년 8월 19일 게재확정일 : 2021년 9월 14일
원고유형 : 학술대회 Fast-Track 교신저자 : 김우주

1. 서론

4차 산업혁명 시대에 들어서며, 급격한 기술 발전과 함께 산업 구조가 다변화되고 있다. 이런 변화에 발맞춰, 대학생들은 학부에서부터 진로를 탐색하고, 희망하는 진로에 도움되는 역량을 쌓기 위해 노력하고 있다. 역량을 키우는 방법으로는 자격증 취득, 대외활동, 기업 인턴쉽 등 다양한 방법이 있으나, 가장 기초적인 것은 전공수업 수강을 통한 지식함양이다(Eun-Seok Kim,

* 이 논문은 국토교통부의 스마트시티 혁신인재육성사업으로 지원되었습니다.

2015). 따라서 학생들이 희망하는 진로와 관련된 전공 수업이 어떤 것인지 파악할 수 있도록 돕는 것이 대학 교육의 중요한 과제라고 할 수 있다. 그러나 대학 전공과목의 수와 다양성이 점차 증대됨에 따라 학생들은 자신에게 맞는 전공과목을 선택하는 것에 어려움을 겪고 있다(Koo, et al., 2019). 일반적으로 학생들은 동기들의 선택이나 선배들의 조언과 같은 경험에 의존하여 수업을 선택하는데(Park, 2001), 이는 일반적인 상황을 고려할 수 있다는 장점이 있지만, 개인의 성향과 기존 수강 과목에 대한 고려가 반영되지 않으며 특정 학생들 간에만 정보가 공유되는 정보의 불평등을 초래하는 문제점이 있다. 또한 최근에는 비대면 수업이 시행되며 학생들 사이의 교류가 적어지면서 경험 기반의 의사결정 마저도 잘 이루어지지 않고 있다(Hayeon Lee et al., 2021). 따라서 본 연구에서는 개인의 특성에 맞는 대학 전공수업을 경험이 아닌 데이터에 기반하여 추천할 수 있는 추천 시스템 모델을 제안하고자 한다.

추천 시스템은 특정 사용자가 관심을 가질 만한 정보, 컨텐츠 (음악, 영화, 책, 이미지 등)를 추천하는 것이다. 이미 유튜브, 페이스북 등과 같이 개인의 성향을 고려하는 것이 중요한 서비스에서 많이 활용되고 있으며, OTT(Over-the-top media service)와 같은 컨텐츠 서비스에서 개인 맞춤형 서비스를 제공하는 것에서 친숙하게 경험할 수 있다. 수업도 정해진 컨텐츠 목록에서 개인에 맞는 수업을 선택한다는 측면에서 일종의 컨텐츠 소비이다. 하지만 다른 컨텐츠 소비와는 다르게 선택의 결과가 초래하는 영향도가 크다는 특징이 있다. 예를 들어 음악과 영화의 경우, 보통 1회성으로 소모되며 컨텐츠를 소비하는데 요구되는 시간이 짧다. 그렇기에 각각의 아이템이 갖는 중요도가 상대적으로 낮으며 선택하는데 있어 깊은 고민이 수반되지 않는다. 전공수업은 보통 1학기동안 수강해야하기에 소비 요구 시간이 길고, 선택된 수업의 구성에 따라 진로, 졸업요건 등 많은 것에 영향을 미치기 때문에 각각의 아이템이 갖는 중요도가 높으며 선택에 있어 보다 큰 신중함이 요구된다. 이런 전공수업이라는 아이템이 갖는 고유한 특성에 따라 교육 분야에서 추천 시스템은 다른 분야보다 상대적으로 적은 수의 아이템 범위를 가짐에도 불구하고 아이템 추천이 의미가 있으며 경험 기반 의사결정에서 반영하지 못하는 개인의 특성을 반영한 의사결정을 지원한다. 실제로 추천 시스템을 활용한 맞춤형 교육에 대한 연구가 국내에서 진행된 바 있다(Gui-Jung Kim et al., 2010). 이러한 연구 동향에 발맞춰 맞춤형 교육 지원 시스템을 구축함으로써 대학 교육의 질을 향상시키고, 학생들의 교육 만족도를 제고하는 것은 필수적이다.

본 연구에 사용된 데이터는 대학교 학부생들의 2015년~2017년 수강 이력 데이터로, 학생들의 개인정보를 비식별화 한 뒤 학생들이 어떤 과목을 수강하였는지에 대한 정보만을 활용하였다. 이 정보는 컨텐츠 소비 여부만을 나타낸 암시적 피드백(Implicit feedback)으로, 본 연구에서는 암시적 피드백 데이터를 이용하는 대표적인 모델인 Neural Matrix Factorization (NeuMF) (He, Xiangnan et al., 2017)의 구조를 활용하였다. 그런데 암시적 피드백만을 사용하는 경우에는, 선호도를 나타내는 명시적 피드백(Explicit feedback)을 사용하는 경우와 달리, 개인의 특성을 나타내는 임베딩 벡터의 표현력이 상대적으로 낮다. 따라서 이를 보완하기 위해, 학생과 수업을 노드로 삼아 네트워크를 각각 구축하고, 각 네트워크의

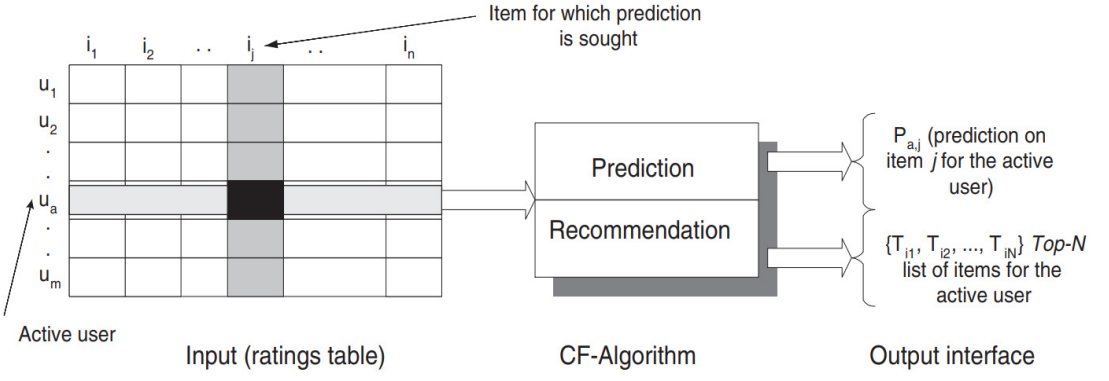
노드를 벡터로 표현하는 방법론인 Node2Vec (Grover and Leskovec, 2016)을 적용하여 높은 표현력을 지닌 벡터를 얻었다. 그리고 이 벡터를 추천 모델의 입력 값으로 활용하여 개선된 성능을 확인하였다.

본 연구는 암시적 피드백이 갖는 제한적 정보라는 한계점을 네트워크 분석을 통한 임베딩으로 극복함으로써, 학생 및 수업에 대해 높은 표현력을 가진 벡터를 추출할 수 있었고, 이를 활용해 좋은 성능을 보이는 전공과목 추천 시스템을 제안하였다. 이는 학생의 수강 이력만으로 전공과목을 추천함으로써 학생들이 기존 경험 기반의 의사결정에서 겪었던 정보의 불평등과 탐색 피로도를 해소시키며 학생들이 진로에 맞는 교육을 받을 수 있도록 돕는다. 또한 이를 통해 학생들에게 시스템적으로 올바른 지도 및 교육 서비스를 제공할 수 있기 때문에 궁극적으로는 대학 교육의 질 향상에 기여한다.

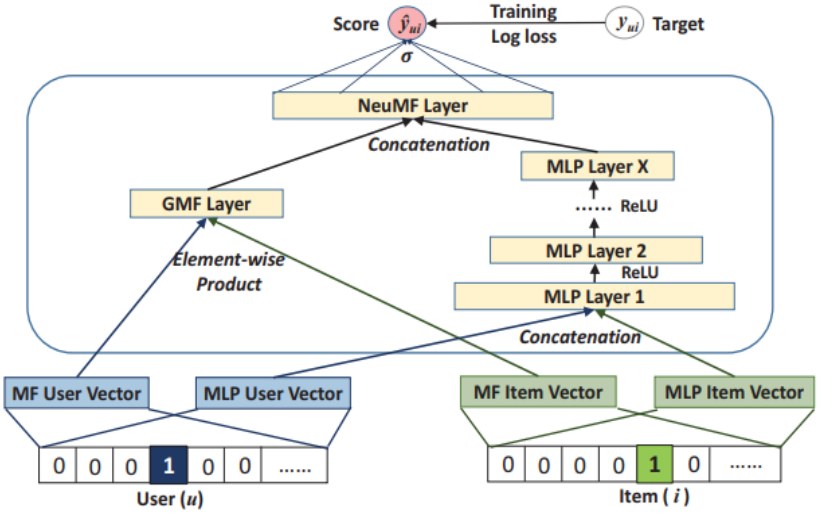
2. 선행 연구

추천 시스템 분야에서는 협업 필터링 (Collaborative Filtering, CF) (Herlocker et al., 2000)이 가장 일반적이며, 최근에는 딥러닝의 발전에 힘입어 딥러닝 기반 추천(Y. Zheng et al., 2016) 등 다양한 기법들이 연구되고 있다.

협업 필터링 : 협업 필터링은 유사성을 바탕으로 비슷한 관심사를 가진 사용자 혹은 특성을 가진 아이템을 예측하는 기법이다. 기본적으로 사용자들의 과거의 경향이 미래에서도 그대로 유지될 것이라 가정하며, 오랫동안 많이 사용되어 온 모델로 성능이 보장된 모델이라 할 수 있다. 협업필터링에는 유사한 사용자를 찾는 유저 기반 관점과 유사한 아이템을 찾는 아이템 기반 관점(Item-based CF) (Sarwar, Badrul, et al., 2001)이 존재한다. 이 중 가장 보편적인 아이템 기반 협업 필터링은 사용자가 선호도를 입력한 기존의 아이템들과 예측하려는 상품과의 유사도를 계산하여 사용자의 선호도를 예측한다. 국내 교육 분야에서는 협업 필터링을 활용해, 학습자 프로파일과 학습 주제 사이의 연관성을 이용해 적합한



〈Figure 1〉 Collaborative Filtering Process



〈Figure 2〉 Neural Matrix Factorization Model

학습을 추천하거나(Ji Won Han et al., 2015), 학습자의 수강 이력 데이터와 수업 정보 데이터를 이용해 교양수업을 추천해주는 연구 (Du Hyeong Kim et al., 2020) 등이 존재한다.

딥러닝 기반 추천 : 딥러닝 기반 추천 시스템인 NeuMF은 협업 필터링의 유저-아이템 상호작용 특성(User-Item interaction feature)에 대한 행렬 분해 (Matrix factorization) (Yehuda Koren et al., 2009)를 학습 가능한 다층 퍼셉트론 (Multi-layer perceptron) 신경망 구조로 대체하여 성능을 개선하였다. 국내 교육 분야에서는 딥러닝 기반 추천을 사용한 연구를 찾아보기 어려웠다.

협업 필터링은 사용자와 아이템의 유사도를 검색하는 과정에서 거리 공식에 의존하는 한계가 있다. 각 사용자와 아이템 간의 유사도 비교를 진행할 때, 모든 차이를 동등하게 반영하기 때문에 구별되는 특성을 유연하게 반영하지 못한다. 딥러닝 기반 추천은 주로 암시적 피드백 데이터를 사용하는 특성상, 개별 데이터의 임베딩 벡터의 표현력이 낮다는 단점이 있다. 임베딩 벡터의 표현력이 낮을 시, 개별 데이터 간의 관계가 정확하게 표현되지 않기 때문에 추천의 정확도가 낮아진다(Zhang, Fuzheng, et al., 2016).

본 연구에 활용함에 있어, 협업 필터링은 수업과 학생에 대한 특성의 차이를 유연하게 반영하기 어렵기 때문에 적합하지 않다. 명시적 피드백 정보인 수업에 대한 학생들의 선호도 데이터가 필요한 아이템 기반 협업 필터링은, 개인 정보가 내포된 교육 분야의 데이터를 이용하는 것에 제약이 있기에 활용되기 어렵다. 또한 단순히 암시적 피드백 정보만을 활용하는 딥러닝 기반 추천 방법론을 쓰는 것도 한계가 있다. 따라서 본 연구에서는 이를 네트워크 분석을 통해 도출한 임베딩 벡터를 NeuMF의 입력으로 활용해 해결했다. 아이템과 유저를 노드로 하며 노드간의 연결(엣지)을 유저와 아이템을 이용해 관계를 정의한 두 개의 네트워크를 생성하였고, 각 네트워크에

대해 노드 임베딩을 통해 유저와 아이템 각각에 대해 표현력이 높은 임베딩 벡터를 추출하여 활용했다. 노드 임베딩의 방법론으로는 무작위 행보(Random walk) 알고리즘을 통해 그래프 구조 정보와 노드 간의 연결 관계를 반영해 벡터를 추출하는 Node2Vec을 활용했다.

3. 전공과목 추천 방법론

본 장에서는 학생 개인에게 적합한 전공과목을 추천하기 위한 방법론을 제시한다. 이 때, 수강기록에 대한 데이터는 비식별화하여 활용하였다. 이는 으로 개인정보에 대한 문제가 발생할 수 있기 때문이다. 수강기록은 학생을 기준으로 수강한 수업과 연도(학기)가 기록되어 있다. 또한, 비식별화 되어있는 학생들의 고유번호와 과목번호, 학생과 수업에 해당하는 전공 명이 메타데이터로 포함되어 있다.

본 장의 구성은 다음과 같다. 제3.1절에서는 제안하는 모델인 Net-NeuMF에 활용하기 적합한 학생 및 수업을 선정하는 기준에 대해 설명하고 조건에 부합하는 학생과 수업에 대해 분석할 것이다. 제3.2절에서는 3.1절의 조건에 해당하는 학생과 수업에 대해 진행한 네트워크 분석에 대해 설명한다. 네트워크 분석은 학생과 수업을 각각 노드로 설정하여 진행하였고 노드의 특성을 나타내는 벡터를 Node2Vec을 이용하여 추출한다. 마지막으로 제3.3절에서는 본 연구에서 제안하는 모델인 Net-NeuMF에 대해 설명한다. 3.1절의 수강기록 데이터에서는 주어지지 않은 부정 상호작용(Negative feedback)을 학습 데이터에 생성하는 과정, 모델의 구조와 학습 및 추천 과정에 대한 설명을 포함한다.

3.1. 학생 및 수업 데이터 전처리

3.1.1. 학생 및 수업 선정 기준

본 연구에 사용한 데이터는 2015년 1학기에서 2017년 2학기까지 1,920,266개의 학생-수업 간 상호작용을 포함하고 있고, 학생의 전공 분류는 총 195개, 수업의 전공 분류는 총 406개이다. 대학 교육의 수강 기록을 나타내는 이 데이터에서는 학원이나 초중고 교육과정과는 다른 특성이 있다. 학생의 입장에서는 휴학으로 인하여 연속적으로 수강을 하지 않는 경우가 있을 수 있고, 특별한 혹은 정형화되지 않은 수강 형태를 가지고 있는 특수 대학(국제학과, 의예과 등), 대학원 등이 있다. 또한 수업은 전 학생들이 공통으로 들어야하는 공통 과목, 필수 과목이 존재한다. 이러한 상황을 모두 고려하는 것은 모델 학습에 있어서 잡음(Noise)로 작용할 수 있기에 학생과 수업에 대해 아래 조건을 적용하여 전처리를 진행했다.

- 2015년 1학기에서 2017년 1학기까지 수강 이력이 있고, 2017년 2학기에 수강한 수업이 한 개 이상인 학생
- 전 학생들이 공통으로 수강해야 하는 공통 과목 및 필수 교양 수업은 제외
- 수강 형태가 정형화되지 않은 전공을 가진 학생 및 수업은 제외

〈Table 1〉 Number of Data after preprocessing

	Number of data	Number of majors
Users(students)	12,031	66
Items(classes)	2,862	144
Interaction	285,482	-

위 세 가지 조건을 적용한 결과는 <Table 1>과 같다. 결과적으로 2015년 1학기부터 2017년 2학기까지 12,031명의 학생이 수강한 수업은 2,862개이고 그 내역(상호작용)은 285,482개이다.

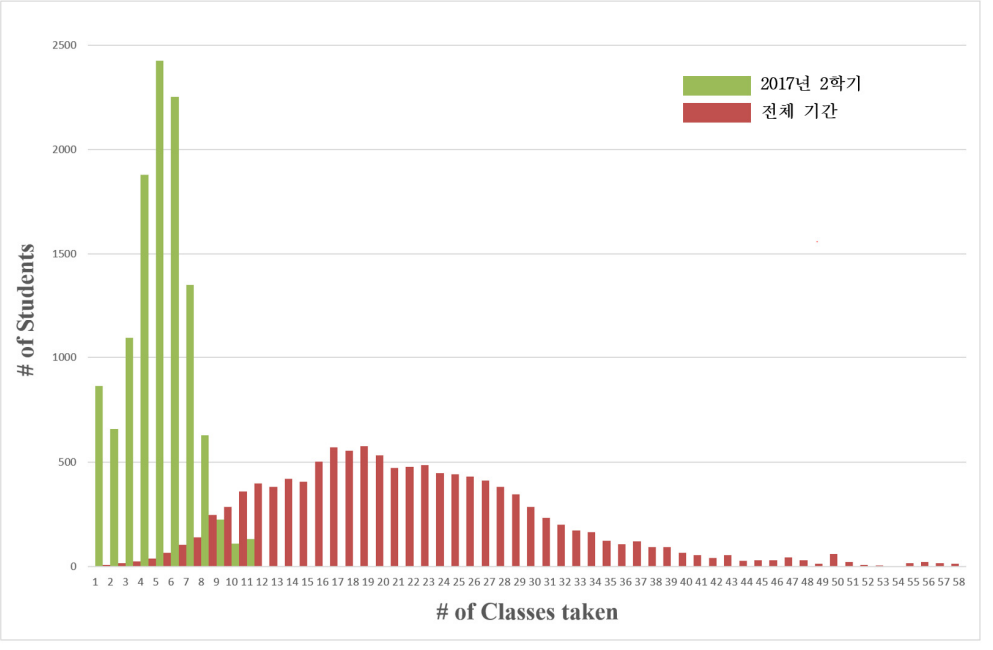
첫 번째 조건은 추천 시스템에서 발생하는 콜드 스타트(Cold-Start) 문제를 완화하고자 적용되었다. 콜드 스타트란, 추천 시스템이 새로운 사용자 또는 아이템들에 대해 충분한 정보가 수집되지 못해 사용자에게 적절한 아이템을 추천해주지 못하는 문제를 말한다. 본 연구에서는 특정 학생이 수강 이력 데이터가 없을 경우 콜드 스타트 문제가 발생한다. 이를 방지하기 위해 활용데이터에서 2015년 1학기에서 2017년 1학기까지 수강 이력이 존재하지 않는 학생을 제외하고 추후 성능평가를 위해 2017년 2학기에 수강한 수업이 없는 학생 또한 제외했다.

마지막 조건에 해당하는 전공의 예시로는 의예과, 간호학과 등이 있다. 해당 전공들에서는 학생들이 선택하는 수업에 있어 차이가 크지 않기 때문에 전공과목 추천의 의미가 제한된다. 해당 조건에 따라 선정한 수업과 제외한 수업 간에는 서로 교차가 적기에 큰 연관성이 존재하지 않아 편향이 크게 발생하지 않을 것으로 판단된다.

3.1.2. 활용 데이터 분석

본 연구에 활용하는 285,482건의 수강 기록(상호작용)에 대한 데이터 확인을 위해 두 가지 분석을 진행하였다.

첫 번째는 수강한 수업의 수에 따른 학생의 수 분포이다 (<Figure 3>). 녹색과 적색의 분포표는



〈Figure 3〉 Distribution with the number of students according to the number of classes taken

각각 특정 기간을 나타내며 X축은 수강한 수업의 수로 특정기간동안 학생이 몇 개의 수업을 수강했는지를 나타낸다. 그리고 Y축은 학생의 수로 해당 개수의 수업을 수강한 학생의 수가 몇 명인지를 나타낸다. (e.g. 적색 분포표의 수강한 수업의 수(X축)가 16인 경우 학생의 수(Y축)이 약 500인 것은 2015년 1학기에서 2017년 2학기까지 16개의 수업을 들은 학생이 약 500명이라는 의미이다.)

적색 분포표는 전체 기간 동안 학생들이 수강한 수업의 수를 확인할 수 있는 데이터로 학습과 평가 데이터셋 분할에 대한 적절성을 파악할 수 있다. 결과적으로 12,031명의 학생이 적당한 수의 학습 데이터셋을 가진다는 것을 확인할 수 있다. 녹색 분포표는 모델 학습 이후 평가를 진행하는데 충분한 데이터셋을 가지고 있는지 확인하는 목적을 가지고 있다. 2017년 2학기에 학생들이 수강한 데이터는 이후 실험 부분에서 평가용 데이터셋으로 사용된다. 하지만, 분포표에서 평가용 데이터셋에서는 학생들이 1개부터 11개까지 다양한 개수의 수업을 수강한 것을 확인할 수 있다. 이는 평가지표를 적용할 때 정밀도(Precision)와 같이 분모에 추천한 개수가 들어가서, 다양한 개수의 수업을 들은 학생의 분모가 같아진다면 모두 동등한 평가가 진행되지 않을 수 있다.

〈Table 2〉 Density of Dataset

	Users	Items	Inter	Dens
ML(1M)	6,040	3,706	1000k	4.47
Abook	70,679	4,915	847k	0.05
Last FM	23,566	48,123	3034k	0.27
Ours	12,031	2,862	285k	0.83

* Inter : Interactions, Dens : Density (%)

두 번째는 <Table 2>에서 확인할 수 있는 데이터 밀도 분석표이다. 이는 전체 데이터에서 상호작용의 비율을 파악해 사용자와 아이템의 관계를 파악하는 것이 의미가 있는지, 학습 모델에 적용할 정도의 데이터를 보유하고 있는지 확인하기 위해 분석을 진행하였다. 데이터의 밀도는 행이 사용자, 열이 아이템으로 구성된 행렬이 있을 때, 각 상호작용이 행렬의 값으로 들어가고 전체 행렬의 값이 얼마나 완전한 정도를 나타내는 지표이다.

표에서 분석한 다른 데이터셋은 추천 시스템에서 사용되는 벤치마크 데이터셋이다. 영화(MovieLens-1m) (Harper, F et al., 2015), 책(Amazon-book) 그리고 음악(Last-FM) (Wang Xiang et al., 2019)으로 구성 되어있고, 각 상호작용에 대한 밀도를 분석했다. 결과적으로 벤치마크 데이터셋과 비교해 사용자, 아이템 수 대비 충분한 상호작용 수를 가지고 있어, 학습 모델에 적용할 수 있을 것으로 판단하였다.

3.2. 네트워크 분석 기반 임베딩 벡터 생성

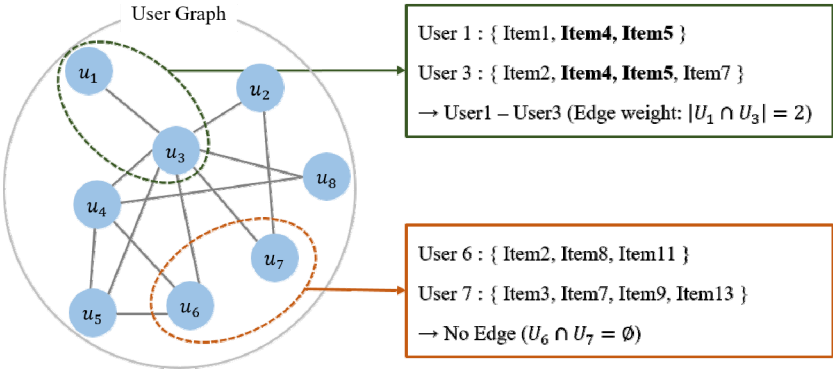
네트워크 분석은 학생과 수업을 잘 표현할 수 있는 벡터(Vector)를 생성하는 것에 목적을 둔다. 기존 모델(NeuMF)은 사용자와 아이템을 원-핫(One-hot) 형식의 벡터로 표현하여 활용한다. 이는 단순 위치만 표현하는 형식으로 개별 사용자와 아이템의 특성을 나타내지 않기에, 그 값을 표현해내는 임베딩(Embedding) 통해 주어지는 벡터의 표현력이 낮다. 따라서, 더 높은 표현력을 가진 벡터로 학생과 과목을 표현하기 위해 네트워크 구축과 Node2Vec을 이용하여 노드 임베딩을 진행하였다.

3.2.1. 네트워크 구축

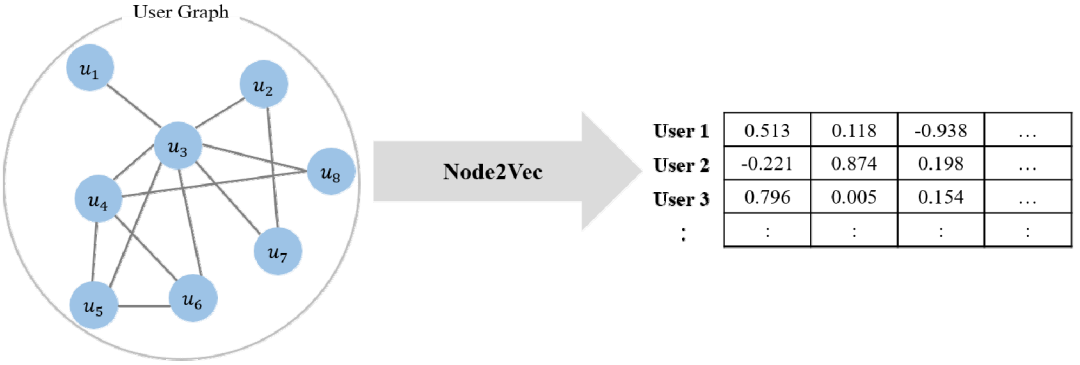
학생의 임베딩 벡터 추출을 위해서, 학생을 노드(Node)로 설정하고 엣지(Edge)는 두 노드의 학생이 같은 수업을 수강한 횟수를 가중치로 하는 유저 네트워크를 구축했다. 이를 통해 학생 간의 관계가 수강 이력 중 겹치는 수업의 수로 표현이 가능하다. 마찬가지로, 수업의 임베딩 벡터를 위해서 수업을 노드로, 두 수업을 공통으로 수강한 학생의 수를 가중치로 하여 수업 네트워크를 구축했다. 이를 통해 수업 간의 관계가 동시에 수강한 학생의 수로 표현이 가능하다. 추가적으로 메타데이터인 학생과 수업의 소속 전공을 속성으로 부여하여 두 네트워크의 표현력을 높였다.

3.2.2. 노드 임베딩

구축한 두 네트워크를 바탕으로 Node2Vec을 활용하여 학생과 수업을 나타내는 임베딩 벡터를 생성했다. Node2Vec은 비유클리디안 공간에 존재하는 그래프의 노드를 유클리디안 공간으로 투영시켜 벡터로 나타내는 표현 학습(Representation learning)인 노드 임베딩 방법론



〈Figure 4〉 Process of constructing network



〈Figure 5〉 Process of node embedding from structured network

중 하나이다. 즉, 앞서 구축한 네트워크의 그래프에서 무작위 행보(Random walk) 알고리즘을 통해 그래프 구조 정보와 노드 간의 연결 관계가 반영된 노드들의 저차원상 벡터 값을 계산하여 (Mikolov, Tomas, et al., 2013) 각 노드(학생, 수업)을 나타내는 벡터로 활용한다.

Node2Vec에서는 강조하고자 하는 노드의 특성에 따라 임베딩 벡터를 추출하는 전략이 2가지가 존재한다. 같은 집단에 속하거나 연결이 유사한 노드끼리 가깝게 임베딩 되게 하여, 노드 간의 유사성을 강조하는 너비 우선 표본추출 전략 (Breadth-First Sampling, BFS) (S. Fortunato, 2010)과, 네트워크 안에서 유사한 구조적 역할을 가지는 노드끼리 가깝게 임베딩 되게 하여, 네트워크의 구조성을 강조하는 깊이 우선 표본추출 전략 (Depth-First Sampling, DFS) (P. D. Hoff et al., 2002) 2가지이다. BFS의 경우 본 연구에서 구축한 네트워크에 대입해 해석하면 같은 학과이거나 같은 수업을 많이 수강한 학생끼리 가깝게 임베딩 된다고 할 수 있다. 추천 시스템의 입력으로 더 적합한 벡터를 추출하기 위해, 유사한 노드를 가깝게 임베딩 되게 하는 BFS 전략에 따라 Node2Vec의 하이퍼 파라미터를 설정했다. 이런 하이퍼 파라미터를 적용하여 도출한 임베딩 벡터가 추구하는 전략에 맞게 표현력이 높은지 확인하기 위해 최근접 이웃 알고리즘(Dudani, Sahibsingh A, 1976)을 이용하여 수업과 학생에 대해 같은 소속 학과 및 단과대학끼리 군집화가 잘 진행되었는지 확인했다. 그 결과 BFS 전략을 택한 Node2Vec이 학생과 수업의 특성을 충분히 표현한다고 판단하여 이를 활용한 임베딩 벡터를 생성하였다.

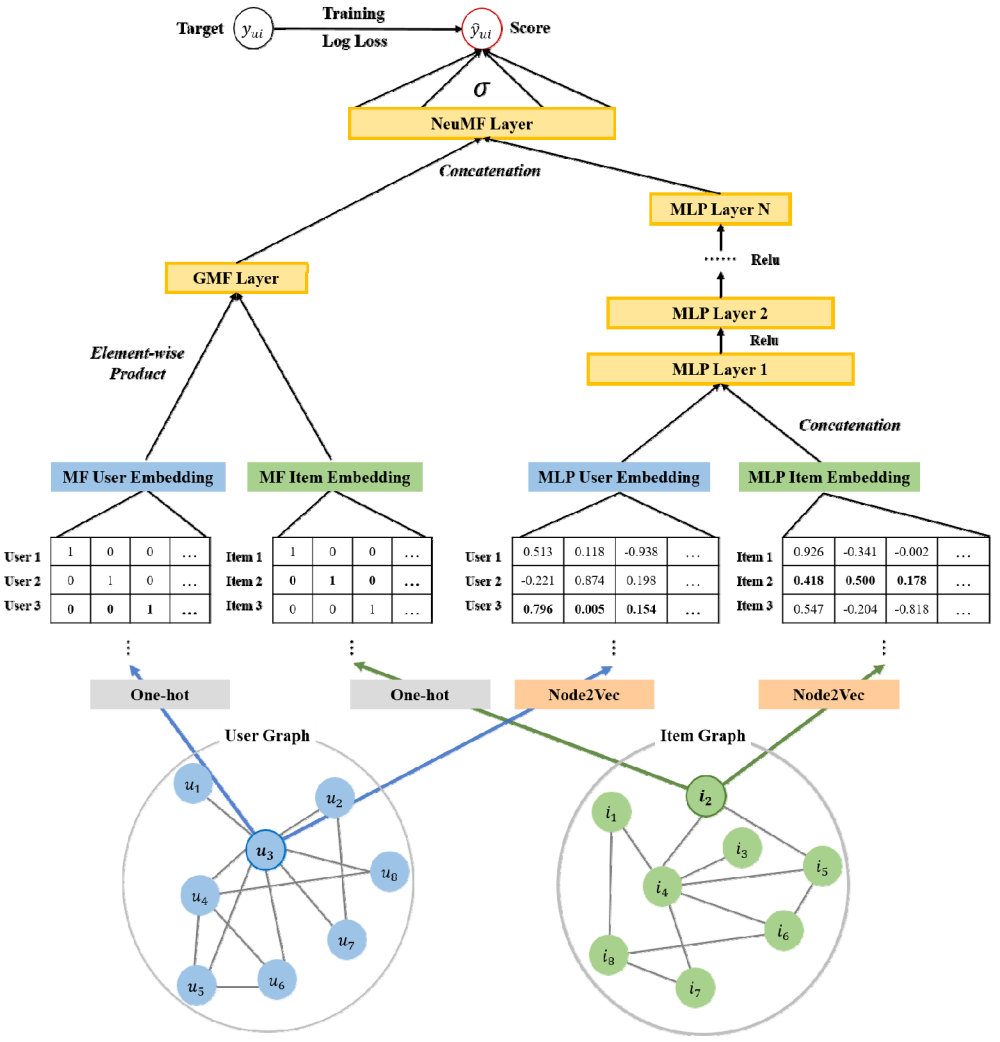
3.3 Net_NeuMF 모델

3.3.1. 학습 데이터 구축

본 연구의 모델인 Net-NeuMF의 학습을 위해서 총 285,482개의 상호작용 중 2015년 1학기에서 2017년 1학기까지의 데이터 220,708개를 학습 데이터셋으로 활용하였다. 그리고 테스트 데이터셋은 2017년 2학기 데이터 64,774개로 학습 데이터셋을 기반으로 학습한 모델을 통해 추천한 결과의 레이블(label)로 평가를 진행한다.

위에서 정한 학습 데이터셋은 학생이 수강한 수업에 대한 정보만을 가지고 표현되어 있다. 하지만, 모델의 학습을 위해서는 부정 상호작용(Negative feedback)이 같이 포함되어 있어야한다. 이에 각 학생은 수강과목에 4배에 해당하는 부정 상호작용의 개수를 포함하여 학습을 진행하고, 부정 상호작용에 해당하는 수업은 학생이 2015년 1학기에서 2017년 1학기 동안 듣지 않은 수업 중 임의로 추출한다. 결과적으로 학습 데이터셋으로 활용하는 긍정 상호작용은 220,708개이고 이들 각각에 4개씩 부정 상호작용을 임의로 추출하기 때문에 총 882,832개의 부정 상호작용을 학습 데이터셋에 포함시켜 총 학습에 활용되는 학습 데이터셋은 1,103,540개이다.

이전 단계에서 얻어진 학생, 수업을 나타내는 임베딩 벡터와 학습 데이터셋은 <Figure 6>과 같이 GMF Layer와 MLP Layer로 각각 입력 값으로 사용된다. GMF 층은 행렬 분해(Matrix Factorization)을 기반으로 하는 모델이다. 이 부분의 입력은 학생과 수업의 위치 정보만을 나타내는 원-핫 벡터가 활용된다. 또한, 기존 행렬 분해 방식과 달리 GMF에서는 각 요소마다의 곱(Element-wise product)을 진행한 이후에 가중치(weight)를 곱하는 방식으로 층이 구성된다. 그리고 MLP 층은



〈Figure 6〉 Structure of Net-NeuMF

다층 퍼셉트론(Multi-layer perceptron) 모델로 총 4개의 층으로 구성되어 있다. 앞서 설명한 GMF 층에서는 원-핫 벡터를 활용했지만, 해당 부분에서는 Node2Vec을 이용하여 생성한 학생, 수업 임베딩 벡터를 입력 값으로 활용한다. 최종적으로 이 두 부분의 결과값을 결합하여 학생이 수업을 선호하는지에 대한 점수를 계산하고 학습 과정에서는 손실 함수(Loss function)로 수업 수강 여부 레이블에 근접하도록 층의 가중치를 업데이트 한다.

4. 실험

4.1. 실험개요

4.1.1 데이터셋 설명

실험에 사용된 데이터는 대학교 학부생들의 2015년부터 2017년까지 3년간의 수강 이력 데이터와 학생과 수업의 기본정보 데이터를 이용했다. 학생의 기본정보로 이름, 생년월일, 주소 등의 개인정보가 포함되지 않은 데이터를 이용했으며 개인의 신상정보는 삭제되고, 학번은 아이디 값으로 대체되어 활용했다. 수업의 기본정보로는 개설년도, 학기, 교과번호, 소속단과대명, 소속학과명을 이용했다. 수강 이력데이터의 경우, 비식별화된 학생별로 2015년도부터 2017년도 사이에 수강한 수업의 목록으로 구성되어 있으며, 이를 학생과 수업의 기본정보 데이터를 연결시켜 모델 실험 및 학습에 이용할 데이터셋을 구축했다. 총 12,031명의 학생과 2,862개의 과목으로 구성되어 있다.

4.1.2. 비교모델 설명 (baseline)

본 연구에서 제안하는 Net-NeuMF 모델의 성능을 평가하기 위해 NeuMF모델의 성능을 비교한다. NeuMF 모델은 사용자의 과거 피드백을 항목 순위에 사용하는 최첨단 딥러닝 기반 추천 시스템이다. MF와 다층 퍼셉트론(MLP) 모델을 결합하고 모델 학습에 부정 표본추출(Negative sampling)을 활용한다.

4.1.3. 평가지표 설명

모델 평가는 추천 정확도를 기준으로 하고 2015년 1학기부터 2017년 1학기까지의 수강 정보를 바탕으로 강의 추천을 진행하여 그 결과를 2017년 2학기 수강 정보와 비교하였다. 조건에 해당되어 학습 및 테스트를 진행한 학생의 수는 12,031명이다. 이 과정에서 사용한 평가지표로는 총 4가지가 있다.

정밀도 (Precision) : 모델이 추천한 수업들 중 실제로 수강한 수업의 비율을 뜻한다.

재현율 (Recall) : 실제 수강한 수업들 중 모델이 추천한 수업의 비율을 뜻한다.

mAP (mean average precision) : 추천 순서에 가중치를 두고 평가하는 방법론이다. 추천 리스트를 통과하며 정밀도를 계산하는 기법으로, 먼저 추천한 수업이 틀렸을 경우 더 높은 점수가 깎이게 되고 나중에 추천한 수업이 틀렸을 경우 더 낮은 점수가 깎이게 된다.

nDCG (Normalized Discounted Cumulative Gain) : 추천의 순서에 더 가중치를 두어 성능을 평가하는 지표이다. 각 아이템들의 선호도(본 연구에서는 평점기반이 아니기 때문에 모두 1로 설정)를

〈Table 3〉 Evaluation metrics (Embedding size = 64, %)

Model	Precision	Recall	mAP	nDCG
NeuMF	24.74	60.58	43.04	56.42
Net-NeuMF (Ours)	25.16	61.21	44.04	56.90
RI	+1.70%	+1.04%	+2.32%	+0.85%

기반으로 관련성이 높은 것을 순서대로 예측하는 평가지표이다.

4.2. Net-NeuMF 성능평가

실험을 하는 과정에서는 학생-수업 상호작용 데이터를 동일하게 사용하였으며, 각각 부정 상호작용 임의 추출 과정이 존재하기 때문에 3회의 반복을 거쳐 평가지표의 평균값을 계산했다. 초기 실험 결과는 NeuMF와 Net-NeuMF의 임베딩을 하는 과정에서 64차원으로 고정하여 진행했다. 또한, Net-NeuMF는 Node2Vec의 임베딩 중 64차원에 해당하는 벡터를 생성해 다층 퍼셉트론의 입력 벡터로 사용했다. 그리고 모델의 출력 중 추천하는 개수는 본 연구의 모든 실험에서 10개로 진행하였다.

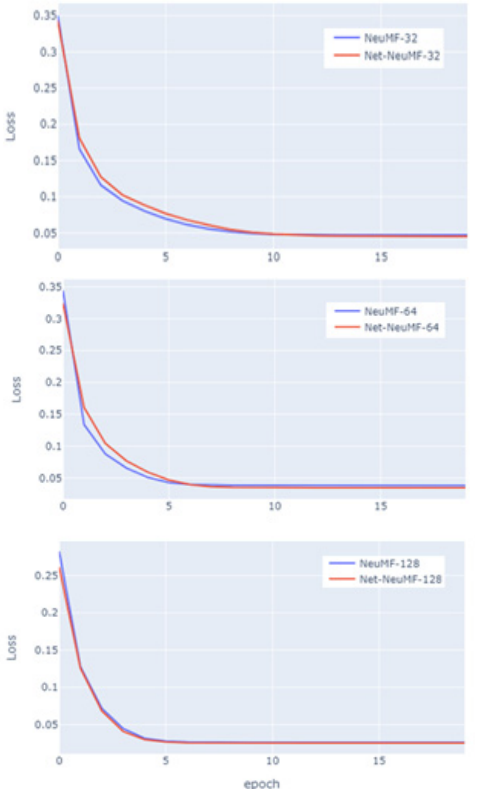
결과적으로 Node2Vec의 64차원 임베딩을 사용했을 때, 기존 NeuMF 구조에서 원-핫 벡터를 사용한 모델보다 정밀도가 1.70%, 재현율이 1.04%, mAP가 2.32%, nDCG가 0.85% 증가한 것을 확인할 수 있다.

정밀도를 계산할 때, 분모는 추천한 수업의 수(본 연구에서는 10개로 고정)로 일정하다. 따라서 학생이 한 과목을 들었을 경우 정확히 추천 시 100%, 잘못 추천 시 0%의 결과가 도출된다. 재현율을 계산할 때는 학생이 실제 수강한 수업의 수가 분모로 활용된다. 이에 따라 개인마다 평가지표 변동성이 큰 정밀도는 상대적으로 높은 개선도를, 변동성이 적은 재현율은 상대적으로 낮은 개선도를 보인다고 판단된다.

mAP와 nDCG는 모두 추천한 과목의 순서에 기반을 두고 성능을 평가한다. 하지만 두 지표에서 개선 정도의 차이가 나는 이유는 nDCG의 아이템 선호도에 의한 것으로 판단된다. 일반적으로 선호도는 데이터의 평점 값이 많이 사용되는데, 본 연구에서는 관련 평점을 책정할 수 없어 모두 1로 통일했다. 이에 개인 선호도가 반영되지 않았고, nDCG에 영향을 미치는 한 변수를 제약한 것이기에 개선도가 크지 않은 것으로 판단된다.

4.2.1. 학습 손실 값 비교

이후 학습 과정에서 사용하는 손실 함수의 값 분포를 확인해 학습 정도를 비교하고, 하이퍼 파라미터인 임베딩 차원의 크기에 대한 실험을 이어서 진행했다.



〈Figure 7〉 Comparison of Training Loss

Net-NeuMF 모델을 학습하는 과정에서 레이블과 모델의 결과 값 차이를 계산하는 손실 함수(Loss function)의 변화에 대해서 <Figure 7>의 그래프를 통해 나타냈다. 각각의 수렴하는 손실 값은 Net-NeuMF(32, 64, 128 차원) 기준 0.045, 0.035, 0.026으로 적절한 학습이 진행되었다. 또한, NeuMF(32, 64, 128 차원)는 0.047, 0.038, 0.026로 손실 값이 수렴했다. 최종적으로 수렴하는 손실 값이 작아질수록 더 좋은 성능을 보이는 것을 확인할 수 있었다

4.2.2. Node2Vec 임베딩 차원 영향도 분석

임베딩 차원은 학생과 수업을 나타내는 벡터의 표현력에 직접적인 영향을 미친다. 따라서, 일반적으로 차원의 크기가 커질수록 더 높은 표현력을 가지기에 좋은 성능을 보인다. 비교 모델인 NeuMF와 본 연구의 모델인 Net-NeuMF 모두 128차원으로 커질수록 좋은 성능을 보이는 경향성을 보였다. 또한, 두 모델에 대해 재현율의 상대적인 증가량을 비교해보면, 32차원은 +0.5%, 64차원은 +1.04% 그리고 128차원은 +0.2%가 증가했다.

임베딩의 목적은 고차원 벡터를 학생과 수업의 관계를 가장 잘 드러낼 수 있는 특성 몇 가지로 차원을 축소하여 불필요한 특성은 제거하고, 모델의 목적에 부합한 특성만으로 재구성해 벡터의 표현력을 높이는 것이다. 이런 임베딩을 통해 모델 입력 벡터의 표현력이 높아질수록 모델의 성능이 더 개선되는 결과를 보인다. 이에 따라 결과를 해석하면 128차원의 경우, 입력 벡터에 본 연구의 모델에 대해 불필요한 정보를 담게 되면서 상대적으로 노이즈를 발생했다고 보인다. 32차원의 경우, 입력 벡터에 본 연구의 모델에 대해 필요한 정보들에서도 일부 손실 또는 압축이 되어 상대적으로 성능이 낮게 향상된 것으로 판단된다.

5. 결론

본 연구에서는 대학 교육에서 전공 수업의 다양화로 인하여, 전공 수업 선택에 어려움을 겪고 있는 학생들의 어려움을 해결하고자 하였다. 연구에 활용한 데이터는 대학교 학생들의 수강 내역과 그 메타데이터이다. 제안하는 모델의 구조는 암시적 상호작용에 활용하는 추천 시스템 모

〈Table 4〉 Impact of Embedding Size (%)

Embedding	Model	Precision	Recall	mAP	nDCG
32	NeuMF	24.52	59.91	42.15	55.23
	Net-NeuMF	**24.59**	**60.23**	**42.56**	**55.71**
64	NeuMF	24.74	60.58	43.04	56.42
	Net-NeuMF	**25.16**	**61.21**	**44.04**	**56.90**
128	NeuMF	25.04	61.31	44.87	57.43
	Net-NeuMF	**25.10**	**61.44**	**44.95**	**57.85**

델인 NeuMF의 구조를 활용하였다. NeuMF의 경우에는 사용자와 아이템의 임베딩으로 원-핫 벡터를 사용하여 각각의 특성을 정확하게 반영하지 못한다는 문제점이 있었다. 따라서 제안하는 Net-NeuMF 모델에서는, 사용자와 아이템의 관계를 통해 정의된 학생 및 수업 네트워크를 구축하고 Node2Vec을 활용한 노드임베딩을 통해 학생과 수업의 특성을 더 효과적으로 표현한 벡터를 생성하여 이를 임베딩으로 활용했다. 이런 네트워크 분석 과정을 통해 암시적 피드백만을 사용하면서도 사용자와 아이템의 특성을 나타내는 임베딩 벡터의 표현력을 높일 수 있었다. 모델 세부구조에서 GMF 층의 입력으로는 기존의 원-핫 벡터를 입력함으로써 기존 NeuMF가 갖춘 장점을 유지하면서도, Net-NeuMF의 다층 퍼셉트론의 입력은 Node2Vec으로 대체함으로써 임베딩 벡터의 표현력을 높여 학습능력이 개선되는 효과를 얻을 수 있었다. 결과적으로 원-핫 벡터만을 활용하는 기존 방법론에 비해 추천의 정확도를 평가하는 여러 지표에서 좋은 성능을 보였다. 특히 모든 학생에게 동등한 평가 방식을 적용한 주요한 의미를 갖는 평가지표인 재현율에서 가장 높은 성능을 보였다. 추가적으로 임베딩 차원에 따른 실험을 진행하여 3가지 차원에서 기존 방법론에 비해 모두 성능이 개선되는 것을 확인했으며, 64차원 임베딩에서 상대적 표현력의 차이가 최대화되는 결과를 얻었다.

본 연구는 연도별로 신설되는 수업으로 인해 발생하는 콜드 스타트 문제를 해결하지 못한 한계가 있다. 이는 수업의 관련정보들을 수업 네트워크의 추가적인 노드 속성으로 부여해 네트워크의 표현력을 높인다면 개선할 수 있을 것으로 생각되며 추후 연구로 진행할 예정이다. 본 연구를 통해 대학 교육 분야에서 개인의 특성을 반영하여 학생에게 맞는 전공과목을 선택하는 것에 기여할 것으로 생각한다.

참고문헌(References)

Dudani, Sahibsingh A. "The distance-weighted k-nearest-neighbor rule." *IEEE Transactions on Systems, Man, and Cybernetics 4* (1976): 325-327.

Grover, Aditya, and Jure Leskovec. "node2vec: Scalable feature learning for networks." *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. (2016)

Han. Ji Won, HS. Lim, “The Recommendation System based on Collaborative Filtering for Adaptive Learning”, *The Korean Association Of Computer Education a collection of essays* 157-160, (2015)

Harper, F. Maxwell, and Joseph A. Konstan. "The movielens datasets: History and context.", *Acm transactions on interactive intelligent systems (tiis)* 5.4 1-19 (2015)

He, Xiangnan, et al., "Neural collaborative filtering.", *Proceedings of the 26th international conference on world wide web*. (2017)

Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. "Explaining collaborative filtering recommendations.", *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. (2000)

Kim Du hyeung, WS. Shin, KW. Han, JS. Lee, KB. Moon, SG. Lee, SY. Han, HJ. Kwon, SW. Han, “A system for recommending

university liberal arts courses using collaborative filtering", *Korean Institute Of Industrial Engineers Autumn Conference a collection of essays* 2551-2556, (2020)

Kim Eun-Seok, "A Study on the Difference of Job Satisfaction by Means of Major Selection Criteria of College Graduates", *The Journal of Career Education Research* 28(3) 85-101, (2015)

Kim Gui-Jung, BH. Kim, JS. Han, "Customizing Intelligent Recommendation System based on Compound Knowledge", *JOURNAL OF THE KOREA CONTENTS ASSOCIATION* 10(8) 26-31, (2010)

Koo Yoo Young, DH Park, JJ Kim, YH Park, CK Ko, BK Lee, "Meta-analysis of course selection data of the university graduates revealed the problems of course structures", *Korean Journal of General Education* 13(2) 369-396, (2019)

Lee Hayeon, JE. Go, MH. Joo, "Effects of University belonging and College Life Satisfaction on Learning Persistence in Non-face-to-face Learning Environment due to COVID-19 Pandemic", *The Journal of Career Education Research* 34(1) 231-251, (2021)

Mikolov, Tomas, et al., "Efficient estimation of word representations in vector space.", *arXiv* preprint arXiv:1301.3781 (2013).

P. D. Hoff, A. E. Raftery, and M. S. Handcock. "Latent space approaches to social network analysis", *J. of the American Statistical Association*, (2002)

Park Roh-Gook, "A Study on Selection Altenatives of Basic and Major Courses for College Students", *Journal of the Korea Industrial Information Systems Research* 6(1) 48-55, (2001)

S. Fortunato. "Community detection in graphs", *Physics Reports*, 486(3-5):75 − 174, 2010.

Sarwar, Badrul, et al., "Item-based collaborative filtering recommendation algorithms.", *Proceedings of the 10th international conference on World Wide Web.* (2001)

Wang Xiang, H. Xiangnan, Y. Cao, M. Liu and T. Chua, "KGAT: Knowledge Graph Attention Network for Recommendation", *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19),* August 4−8, (2019)

Y. Zheng, B. Tang, W. Ding, and H. Zhou. "A neural autoregressive approach to collaborative filtering", *In ICML*, 764−773, (2016)

Yehuda Koren, R. Bell, C. Volinsky, "Matrix factorization techniques for recommender systems.", *IEEE Computer 42(8)* 30−37, (2009)

Zhang, Fuzheng, et al., "Collaborative knowledge base embedding for recommender systems.", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.* (2016).

Abstract

Major Class Recommendation System based on Deep learning using Network Analysis

Jae Kyu Lee* · Heesung Park* · Wooju Kim**

In university education, the choice of major class plays an important role in students' careers. However, in line with the changes in the industry, the fields of major subjects by department are diversifying and increasing in number in university education. As a result, students have difficulty to choose and take classes according to their career paths. In general, students choose classes based on experiences such as choices of peers or advice from seniors. This has the advantage of being able to take into account the general situation, but it does not reflect individual tendencies and considerations of existing courses, and has a problem that leads to information inequality that is shared only among specific students. In addition, as non-face-to-face classes have recently been conducted and exchanges between students have decreased, even experience-based decisions have not been made as well. Therefore, this study proposes a recommendation system model that can recommend college major classes suitable for individual characteristics based on data rather than experience. The recommendation system recommends information and content (music, movies, books, images, etc.) that a specific user may be interested in. It is already widely used in services where it is important to consider individual tendencies such as YouTube and Facebook, and you can experience it familiarly in providing personalized services in content services such as over-the-top media services (OTT). Classes are also a kind of content consumption in terms of selecting classes suitable for individuals from a set content list. However, unlike other content consumption, it is characterized by a large influence of selection results. For example, in the case of music and movies, it is usually consumed once and the time required to consume content is short. Therefore, the importance of each item is relatively low, and there is no deep concern in selecting. Major classes usually have a long consumption time because they have to be taken for one semester, and each item has a high importance and requires greater caution in choice because it affects many things such as career and

* Department of Industrial Engineering, Yonsei University
** Corresponding Author: Wooju Kim
Department of Industrial Engineering, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2123-7754, E-mail: wkim@yonsei.ac.kr

graduation requirements depending on the composition of the selected classes. Depending on the unique characteristics of these major classes, the recommendation system in the education field supports decision-making that reflects individual characteristics that are meaningful and cannot be reflected in experience-based decision-making, even though it has a relatively small number of item ranges. This study aims to realize personalized education and enhance students' educational satisfaction by presenting a recommendation model for university major class. In the model study, class history data of undergraduate students at University from 2015 to 2017 were used, and students and their major names were used as metadata. The class history data is implicit feedback data that only indicates whether content is consumed, not reflecting preferences for classes. Therefore, when we derive embedding vectors that characterize students and classes, their expressive power is low. With these issues in mind, this study proposes a Net-NeuMF model that generates vectors of students, classes through network analysis and utilizes them as input values of the model. The model was based on the structure of NeuMF using one-hot vectors, a representative model using data with implicit feedback. The input vectors of the model are generated to represent the characteristic of students and classes through network analysis. To generate a vector representing a student, each student is set to a node and the edge is designed to connect with a weight if the two students take the same class. Similarly, to generate a vector representing the class, each class was set as a node, and the edge connected if any students had taken the classes in common. Thus, we utilize Node2Vec, a representation learning methodology that quantifies the characteristics of each node. For the evaluation of the model, we used four indicators that are mainly utilized by recommendation systems, and experiments were conducted on three different dimensions to analyze the impact of embedding dimensions on the model. The results show better performance on evaluation metrics regardless of dimension than when using one-hot vectors in existing NeuMF structures. Thus, this work contributes to a network of students (users) and classes (items) to increase expressiveness over existing one-hot embeddings, to match the characteristics of each structure that constitutes the model, and to show better performance on various kinds of evaluation metrics compared to existing methodologies.

Key Words : Big data in Education, Deep learning, Network analysis, Node embedding, Recommendation system

Received : May 22, 2021 Revised : August 19, 2021 Accepted : September 14, 2021
Corresponding Author : Wooju Kim

저 자 소 개

이 재 규
서울시립대학교 공간정보공학, 빅데이터분석학 학사를 취득하고, 연세대학교 산업공학과에서 석사과정 재학 중이다. 주요 관심 분야는 추천 시스템, 컴퓨터 비전, 딥러닝이다.

박 희 성
연세대학교 산업공학과에 학사를 취득하고, 연세대학교 산업공학과에서 석사과정 재학 중이다. 주요 관심 분야는 추천 시스템, 컴퓨터 비전, 딥러닝이다.

김 우 주
1987년 연세대학교 BBA 과정 학사 학위를 취득하고, 1994년 KAIST 경영과학 박사를 취득하였으며, 현재 연세대학교 산업공학과 교수로 재직 중이다. 관심분야는 시맨틱 웹, 시맨틱 웹 환경의 의사결정지원 시스템, 시맨틱 웹 마이닝, 지식관리 및 인공지능 웹 서비스이다.