Neural Collaborative Filtering 기반의 기업 취업을 위한 과목 추천 시스템

박재헌[1], 이원철[2], 은동진[3], 이수안[4]
[1]강원대학교 소프트웨어 융합 연계전공, [2](주)시즈소프트, [3]강원지역혁신플랫폼 대학교육혁신본부,
[4]세명대학교 컴퓨터학부
now1256@kangwon.ac.kr, 2wonchoel@gmail.com, goto95@kangwon.ac.kr, suanlee@semyung.ac.kr

Subject Recommendation System for Corporate Employment based on Neural Collaborative Filtering

Jae heon Park[1], Woncheol Lee[2], Dong Jin Eun[3], Suan Lee[4]
[1] Kangwon Joint Program of Software Convergence Course, [2]Seeds Soft, [3]Innovative Institute of Education, Gangwon Regional Innovation Platform, [4]School of Computer Science, Semyung University

요 약

여러 산업들이 불확실성의 시대에서 디지털 전환을 통해 생존 전략을 모색하고 있다. 본 논문은 대학교의 진로 상담을 추천시스템으로 전환해 대학교의 디지털 전환의 생존 전략을 모색한다. 또한 빅 데이터 시대의 다양한 데이터를 학생들이 판단하기 어렵기 때문에 학생들에게 도움을 주기 위한 추천 시스템을 구현하고자 한다. Collaborative filtering 기반의 추천시스템을 소개하고 이전에 사용된 Matrix Factorization의 방식에 neural network를 사용한 NCF 모델을 사용하여 추천시스템을 구현한다. 그리고 기존의 학생-과목 추천 방식이 아닌 기업 - 과목의 추천 방식을 사용해 일반화된 기업의 과목을 볼 수 있게 추천의 선택지를 늘렸다. 모델의 테스트를 위해 Hit-rate, Precision, Recall을 사용해 모델을 검증했으며, Hit-rate 0.97218이라는 값을 확인하며 모델에서 기업에 일반화된 과목들을 확인할 수 있었다.

1. 서론

여러 산업들은 불확실성 시대에서 디지털 전환이라는 생존전략을 강구하고 있다. 특히 코로나 시기로 인해 디지털 전환은 더욱 중요하게 떠올랐고, 온라인 플랫폼을 활용한 디지털 산업들이 증가하였다. 이에 매체들은 디지털 전환을 통해 다양한 기술들을 만들려고 노력 중이다. 디지털 전환에서 중요시하는 것은 기술이 소비자가 얼마나 친근하게 다가갈 수 있는지를 말하는 소비자 친화적이라는 키워드이다. 소비자 친화적인 기술은 소비자 개인이 기술의 발전으로 인해 이전과 자신의 경험이 얼마나 달라졌는지로 말할 수 있다. 이 논문은 딥러닝 알고리즘의 추천 시스템을 통해 소비자들에게 친화적으로 다가가고 불확실성 시대에서 생존전략으로 추천 시스템을 사용해 전략적으로 생존할 방법을 재고한다. 특히 코로나시기 학생들이 인터넷 매체와 sns를 활용하여 정보를 얻는 것을 파악하고 상담사가 아닌 시스템적으로 분석해 학생들에게 진로에 맞는 과목을 선택, 수강할 수 있게 하는 프로그램인 추천 시스템을 개발하고자 하였다. 이러한 추천 시스템을 구현하기 위해 user에 높은 rating을 갖는 item을 추천할 수 있는 모델인 Neural collabaoritive filtering(NCF) [1] 모델을 사용한다. 또한 본 논문은 학생에게 과목을 추천하는 기존의 방법론과 다르게 기업에 가기 위해 어떤 과목을 수강하는 것이 유리한지를 추천하는 시스템을 고안해 학생들의 진로 선택에 있어 다양성을 늘려 선택의 폭을 늘리는 것을 고안하고자 이런 연구를 진행하게 되었다.

2. 추천 시스템

기존의 추천 방식은 collaborative filtering[2]을 통해 최근접 이웃기반을 통해 사용자-아이템 행렬에 사용자가 평가하지 않은 아이템을 예측하는 Matrix Factorization[3] 방식을 사용했다. 이는 기존에 sparse 한 행렬을 행렬분해를 통해 분해하고 다시 합치는 과정에서 이전에 평가하지 않았던 아이템에 대한 평점을 확인하는 방식으로 사용했다. 행렬분해의 전통적인 알고리즘으로는 고유값 분해(Eigen Value Descomposition, EVD), 특이값분해((Singular Value Decomposition, SVD) [4] 등을 사용해 왔다. Matrix Factorization(MF)[3]에서는 Interaction을 모델링 하기 위해 사용하는 내적(inner-product)으로 latent feature를 선형적으로 곱하는 단순한 방법을 사용해 매우 효율적이지만, 이는 동시에 내적이 선형적인 관계만을 모델링 할 수 있다는 한계를 가지고 있다.

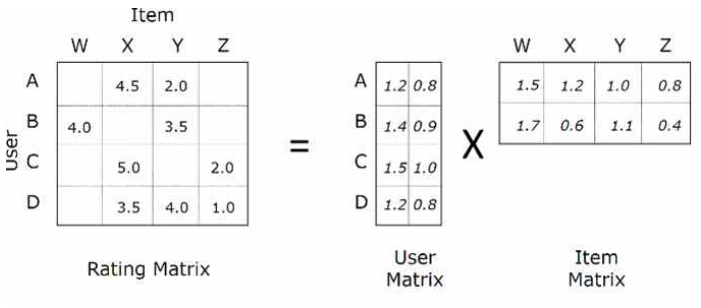

그림 1 Matrix Factorization

NCF(Neural collaborative filtering)[1] 모델은 비선형적인 관계를 모델링할 수 있는 방법인 심층신경망을 통해 데이터 사이의 상호관계를 모델링할 수 있는 모델이다. NCF에서는 GMF와 MLP를 융합한 모델을 제안한다. GMP와 MLP에서는 각각의 Embedding layer를 사용한다. 이 Embedding layer란 input 단계의 sparse 한 벡터를 dense한 벡터로 맵핑하는 단계를 의미한다.

GMF(Generalized Matrix Factorization) 모델은 유저와 아이템의 latent vector 의 element-wise를 구한 값에 활성화 함수 와 latent vector들의 영향력을 조정하는 가중치를 곱하여 유저와 아이템 간의 점수를 구할 수 있다. 이때 활성화 함수에 non-linear의 sigmoid를 적용하여 기존의 linear 방식의 Matrix Factorization 모델보다 더 많은 표현이 가능해진다. Mulit-Layer Perceptron (MLP) 모델은 유저와 아이템 간의 concatenated vector를 여러 hidden layers를 통과시켜 모델에게 높은 유연성과 비선형성을 부여할 수 있다.

MLP는 레이어의 수만큼 활성 함수를 통과시켜 유저와 아이템 간의 점수를 구하게 된다. 최종적으로 user-item 간의 상호관계를 표현하기 위해 MF의 선형성(linearity)과 MLP의 비선형성(non-linearity)을 결합했고, collaborative filtering의 핵심 가치(user와 item의 상호작용 모델링)를 놓치 않으면서 성능을 높였다.

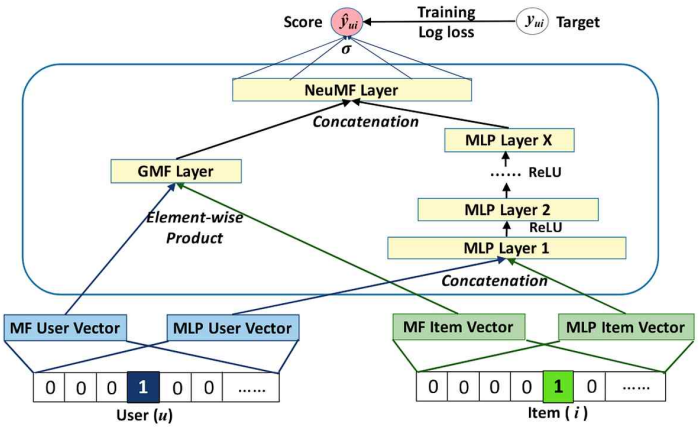

그림 2 Nueral collabaoritive filtering

3. 실험

본 논문에서 사용한 데이터로는 강원대학교 내부의 학생 데이터로 학번, 성별, 단과대, 이수년도, 성적, 과목코드, 취직한 기업, 기업코드로 이루어져 있다. NCF는 implicit data[5]를 사용함으로 데이터에서 기업에 취직한 학생을 대상으로 수강한 과목은 1 수강하지 않은 과목은 0으로 과목을 처리했다. 이때 0으로 표시된 것은 negative한 것이 아닌 단순 듣지 않았다는 것을 의미한다. 총 학생수는 21121명이고 과목코드는 15974개 기업수는 927개로 진행하였다. 원래의 Matrix는 User*item인 학생*과목이지만 현 실험에서는 기업*과목 행렬로 바꿔주었고 그렇게 하기 위해 학생이 취업한 기업을 user로 바꾸고 그 학생이 수강했던 과목들을 기업의 item으로 바꿔서 새로운 기업, item 행렬을 만들었다. 이때 학습용 셋과 평가용셋은 비율은 7:3으로 나누었다.

표 1 강원대 학생 수강 데이터

STID	SX	CCD	RK	YY	HK	GRUP_CD	S_AVG	HCD	HIRED_YY	CO	BZC_CD
2001xxx	2	1105552250.0	1	2001	1	003585	1.5	chj79001	2010.0	강원대학교	85302.0
2001xxx	2	1105552250.0	1	2001	1	100014	3.0	chj79001	2010.0	강원대학교	85302.0
2001xxx	2	1105552250.0	1	2001	1	006808	2.0	chj79001	2010.0	강원대학교	85302.0
2004xxx	2	3201432190.0	1	2001	1	335575	3.5	chj79001	2010.0	강원대학교	85302.0
2004xxx	2	3201432190.0	1	2001	1	009809	4.0	chj79001	2010.0	강원대학교	85302.0
...	...	...	...	...	...	...	...	...	...	...	...

표 2 추천에 사용한 데이터

STID	GRUP_CD	BZC_CD
2001xxx	003585	85302.0
2001xxx	100014	85302.0
2001xxx	006808	85302.0
2004xxx	335575	64552.0
2004xxx	009809	64552.0
...	...	...

NCF 모델의 하이퍼파라미터는 손실함수 Binary Cross Entropy(BCE), 옵티마이저 Adam, 배치사이즈 256, 학습률 1e-4, 에폭 1000으로 하이퍼 파라미터를 적용하였다. 모델의 구조는 GMF와 MLP layer가 concatenate 되어 있으며 GMF와 MLP layer에서는 각각의 embedding layer로 user와 item을 embedding 진행한다. MLP layer에서는 user와 item의 latent vector를 concatenate를 한 vector를 신경망에 넣어 비선형적인 데이터 관계를 학습한다. GMF에서는 이전의 MF에서 user와 item을 임베딩한 것을 element-wise를 user와 item 간의 점수를 구한다.

모델 평가를 위해 추천모델에서 사용하는 지표인 Hit-Rate, Precision, Recall을 사용해 모델을 평가하였다.

기존의 추천시스템은 모든 user를 학습하고 item에 빈칸을 뚫어 실제로 학습했을 때 기존에 있던 item이 추천되는지로 Hit-Rate, Precision[6], Recall[6]을 진행한다. Hit-Rate의 경우 식 (1)과 같이 적중률로서 전체 사용자 수 대비 적중한 사용자 수를 의미한다.

Precision[4]의 경우 정밀도로 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율로 표현하며 식 (2)와 같다. Recall[4]의 경우 재현률로 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율로 표현하며 식 (3)과 같다. 하지만 현 실험에서는 user가 기업이기 때문에 기존의 방식으로 진행하기 어렵다. 그러므로 기업에서 추천하는 과목의 Top-k개를 뽑아 해당 기업에 취직했지만, 학습에 들어가지 않은 테스트 데이터셋에서 학생의 과목에 user가 추천한 과목이 있는지 없는지로 평가를 진행했다.

$$Hit\_Rate@K = \frac{(\text{추천한과목} \cap \text{수강과목})\text{있는 학생수}}{\text{학생의수}} \quad (1)$$

$$\text{Pr}esicion@K = \frac{(\text{학생들이들은과목} \cap \text{추천한과목})\text{의수}}{\text{추천한과목의수}} \quad (2)$$

$$Recall@K = \frac{(\text{학생들이들은과목} \cap \text{추천한과목의})\text{수}}{\text{학생들이들은과목의수}} \quad (3)$$

본 실험의 Test에서는 Top k를 10, 25, 50, 100으로 나눠서 진행했다. Hit-Rate의 경우 Top 10과 Top 100의 경우 0.83992와 0.97218로 준수한 성능을 보여줬다. 기업에 간 학생들의 과목을 item으로 썼기 때문에 기업에 대한 과목이 일반화 되어 성능이 높은 모습을 보여준다. 실제로 그림 3과 같이 기업에서 과목을 추천할 때 대체로 같은 과목들을 추천하는 것을 확인할 수 있다.

```
239: [1407, 5, 642, 368, 238, 329, 367, 697, 698, 53],
675: [1407, 5, 642, 368, 329, 697, 238, 699, 698, 367],
854: [5, 1407, 642, 329, 368, 697, 699, 238, 367, 698],
549: [5, 1407, 642, 329, 368, 699, 697, 698, 367, 238],
430: [5, 1407, 642, 368, 329, 697, 699, 238, 698, 367],
201: [1407, 5, 642, 368, 329, 238, 53, 698, 311, 699],
408: [1407, 5, 642, 368, 329, 238, 698, 53, 367, 311],
```

그림 3 Top-10개의 item 추천

Precision과 Recall의 경우 일반적으로 학생들이 졸업할 때까지 과목을 50과목 정도 듣기 때문에, Top 50에서 각각 0.13600, 0.13185로 비슷한 성능을 보인다.

표 3 NCF 모델 성능 결과

Topk	Hit-Rate	Precision	Recall
10	0.83992	0.19604	0.03758
25	0.93603	0.16723	0.08081
50	0.96260	0.13600	0.13185
100	0.97218	0.09999	0.19377

4. 결론 및 향후 연구

본 논문에서는 학생들이 코로나 시기 이후 인터넷, SNS를 활용해 정보를 얻는 것을 파악하고 고전적인 상담 방식이 아닌 추천알고리즘을 통해 학생 진로 추천 시스템을 고안하였다. 데이터로는 강원대학교 내부 학생의 데이터를 토대로 추천시스템을 고안해 현 강원대학교 학생들이 진로를 찾는 것을 더욱 효과적으로 할 수 있게 만들었다. 이러한 시스템을 만들기 위해 딥러닝의 추천 모델인 collaborative filtering[2]에 neural network를 적용한 NCF[1] 모델을 사용했다. 또한 기존의 학생-과목의 추천 방식과 다르게 기업-과목이라는 학생 추천의 새로운 방향성을 제시했다 본 논문에서는 학생들이 아닌 기업을 대상으로 과목을 추천했기 때문에 기업에 대한 일반화 된 과목이 나온다는 의의가 있다. 현재의 모델에서 데이터셋을 전환하는 과정인, 학생-과목에서 기업-과목의 데이터셋을 처리하는 과정 속에서 interaction이 더욱 잘되는 방법의 연구를 고안해서 모델의 성능을 올릴 예정이고 한 학생의 개인 맞춤 형식의 모델을 고안해 연구를 할 예정이다.

참고 문헌

[1] Xiangnan, Lizi, et al,(2017), Neural Collaborative Filtering, 26th International Conference on World Wide Web ,April , 2017, Pages 173-182

[2] Schafer, Frankowski, et al, (2007), Collaborative Filtering Recommender Systems Part of the Lecture Notes in Computer Science book series (LNISA,volume 4321), 2007, Pages 291-324

[3] Koren, Bell, et al ,(2009), Matrix Factorization Techniques for Recommender Systems, in Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, Pages 30-37

[4] A. Hoecker, V. Kartvelishvili, (1996), SVD approach to data unfolding," Nuclear Instruments and Methods in Physics Research A, 1996.

[5] Bayer, Rendle et al,(2017), A generic coordinate descent framework for learning from implicit feedback. 2017

[6] Jesse Davis , Mark Goadrich (2006), The relationship between Precision-Recall and ROC curves, ICML '06: Proceedings of the 23rd international conference on Machine learning ,June ,2006 , Pages 233-240