

BERT 임베딩을 이용한 개인화 온라인
추천 시스템

Personalized Online Course Recommender
System Based on Bert Embedding

경희대학교 석사 논문 저자 이태영
지도 교수 김재경, 육순형

목차

- **1.서론**

- 1.연구의 배경 및 목적

- **2.관련연구**

- 2.1 CF
- 2.2 NCF
- 2.3 자연어처리
- 2.4 BERT

- **3.텍스트 정보를 반영한 딥러닝 기반의 추천시스템**

- 3.1 BERT기반의 점수 예측 모델(BERT_PM)
- 3.2 강의추천모델

- **4. 데이터 분석 및 실험**

- 4.1 실험 데이터
- 4.2 실험 방법 및 모델 학습
- 4.3 실험 결과

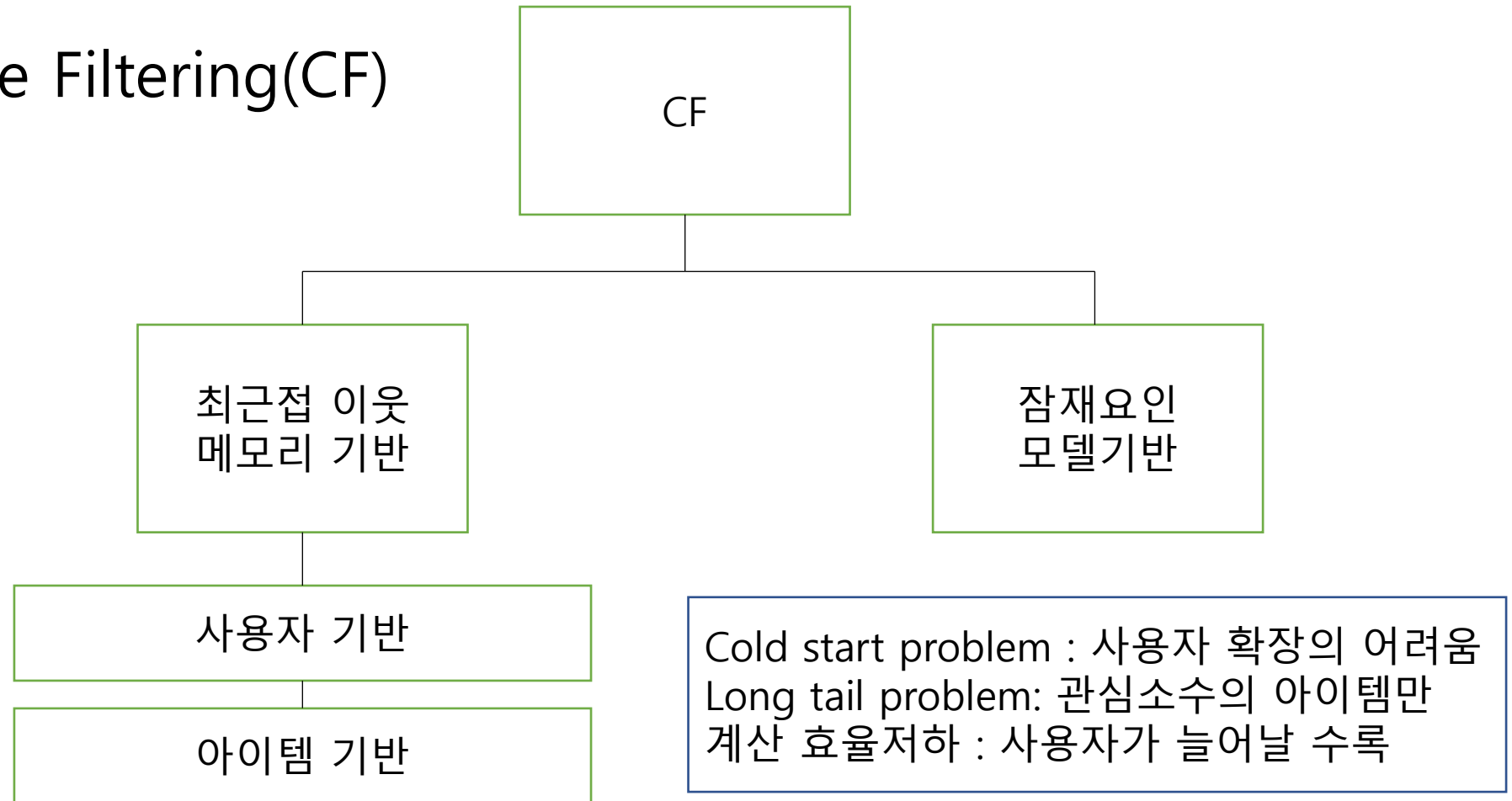
- **5.결론**

1.Abstract

소비자의 구매 의사결정에 영향을 미칠 수 있는 리뷰와 같은 텍스트 데이터를 활용할 수 있도록, NCF(Neural Collaborative Filtering)와 BERT(Bidirectional Encoder Representations from Transformers)를 결합한 딥러닝 기반의 추천 시스템을 제안한다. 제안한 모델은 자연어 처리 딥러닝 모델인 BERT 를 이용하여 사용자 리뷰에서 유용한 특성을 추출하여 사용자 아이템 사이의 고차원적 상호작용을 포착한다

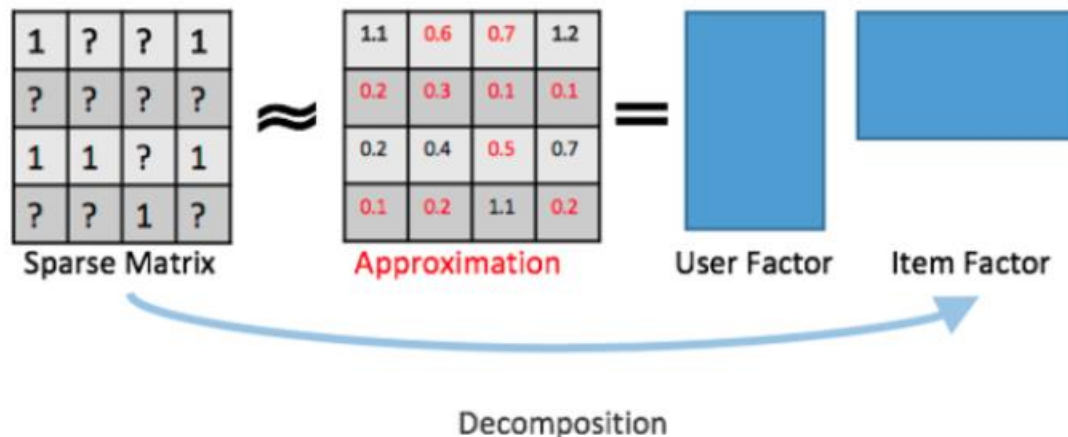
2. 관련 연구

- Collaborative Filtering(CF)



2. 관련 연구

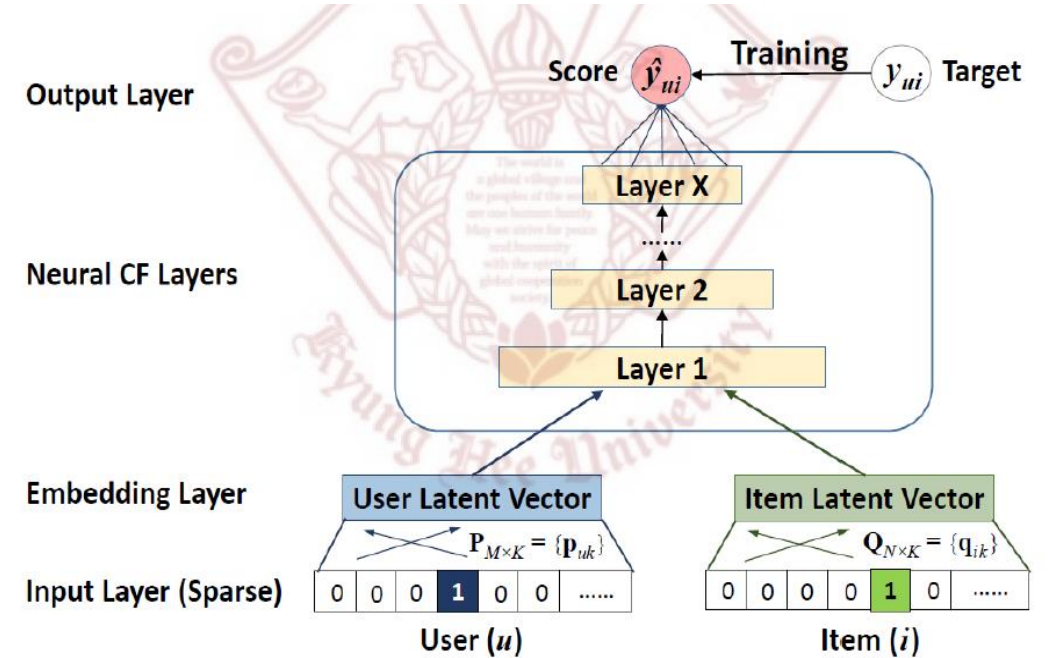
- Collaborative Filtering(CF)
- Latent Factor Model
- Singular Value Decomposition(SVD) 특이 값 분해
- Principle Component Analysis(PVD) 주성분 분석



차원 축소를 통해 최종 적으로 습득한 잠재 특성들의 inner product 로 사용자와 아이템간의 관계를 파악하지만, 사용자와 아이템 사이의 관계를 선형적 방법을 이용하기에 복잡한 관계를 표현하기에 어려움이 있다

2. 관련 연구

- Neural Collaborative Filtering(NCF)



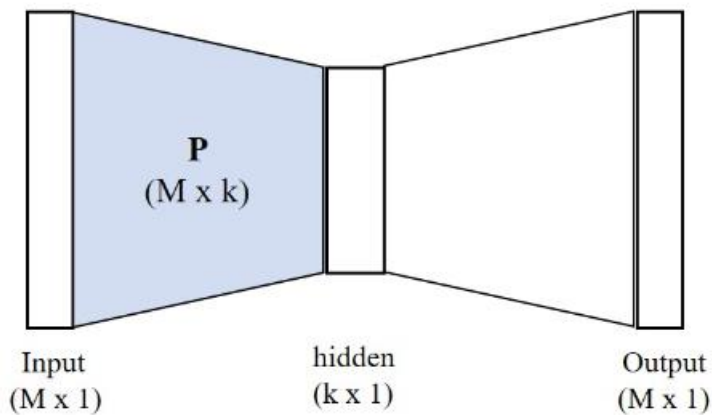
- 행렬 분해법을 인공신경망 형태로 구현한 모델
- 임베딩층에서 입력 단계의 희소 벡터를 밀집벡터로 맵핑 후 유의미한 특성을 찾음
- Implicit feedback: 관계 파악이 어렵지만, 많은 양의 데이터를 구할 수 있기 때문에 신경망 구조의 유효성을 검증하기에 알맞다고 할 수 있음

$$y_{ui} = \begin{cases} 1, & \text{if interaction (user } u, \text{ item } i) \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases}$$

2. 관련 연구

- Neural Collaborative Filtering(NCF)

- Embedding 층에서 fully connected neural network를 사용해 가중치 행렬 P 를 얻게 된다.
- 이 P 행렬의 각 row는 각 user를 표현하는 저차원의 dense 벡터가 되고 이를 user latent vector로 사용하게 된다.



- user latent vector = $P^T v_u^U$ (이때, v_u^U 는 user u 를 나타내는 one-hot 벡터)

- item latent vector = $Q^T v_i^I$ (이때, v_i^I 는 item i 를 나타내는 one-hot 벡터)

Output-layer

$y_{u,i}$ 의 추정치인 $\hat{y}_{u,i}$ 를 구하는 단계다. 여기서 $\hat{y}_{u,i}$ 는 user u 와 item i 가 **얼마나 관련** 있는지를 나타내며 그 값은 0과 1 사이가 된다. 이때 ϕ_{out} 에는 logistic 함수나 probit 함수를 사용한다.

$$\hat{y}_{u,i} = f(P^T v_u^U, Q^T v_i^I | P, Q, \Theta_f) = \phi_{out}(\phi_X(\dots \phi_2(\phi_1(P^T v_u^U, Q^T v_i^I))\dots)), \quad 0 \leq \hat{y}_{u,i} \leq 1$$

2. 관련 연구

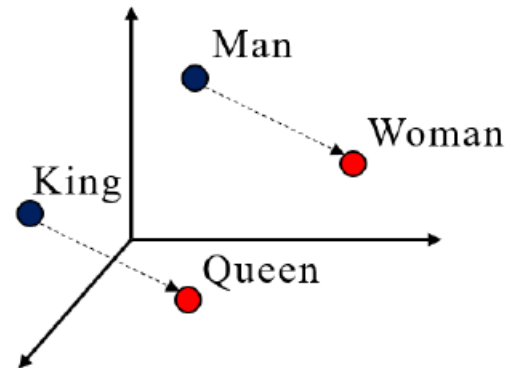
- Natural Language Processing(NLP)

- 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 일을 자연어 처리라고 함
- 원 핫인코딩(one-hot encoding) : 하나의 값만 1이고 나머지는 0으로 표현
- 워드 임베딩(Word embedding): 다차원 공간에 텍스트를 밀집 표현한다. 단어 표현에 있어서 필요 단어의 수보다 적은 차원의 밀집 벡터를 이용하여 학습

Banana
Apple

Apple = [1, 0, 0, 0, 0, ... , 0]
Banana = [0, 1, 0, 0, 0, ... , 0]
Cherry = [0, 0, 1, 0, 0, ... , 0]
Grape = [0, 0, 0, 1, 0, ... , 0]

(a) One hot encoding



(b) Embedding vector

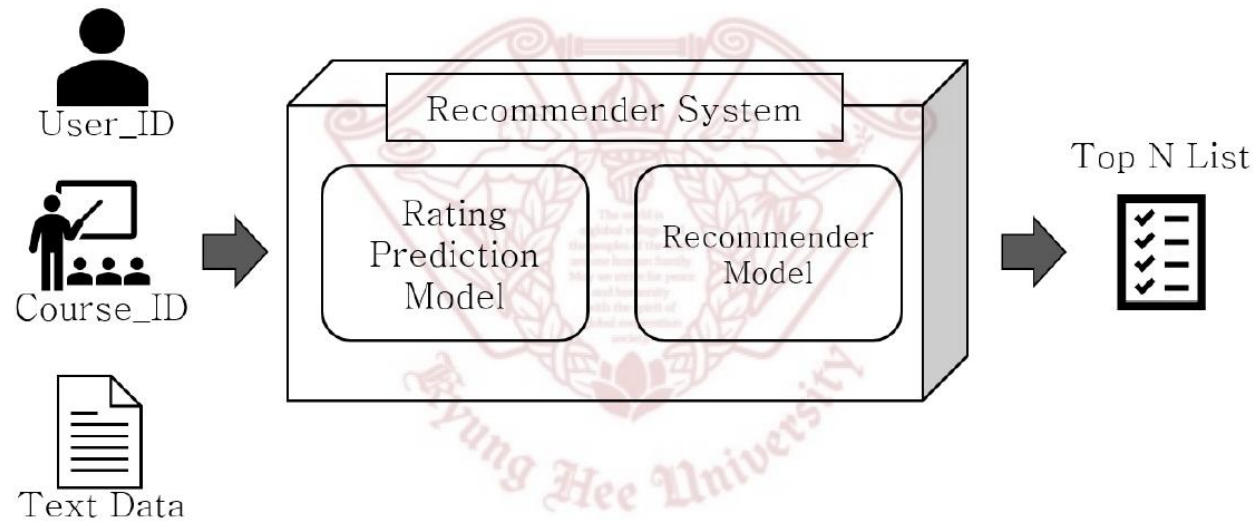
2. 관련 연구

- Bidirectional Encoder Representations from Transformers(BERT)
- BERT 는 2018 년 구글이 공개한 자연어 처리 딥러닝 모델이다. BERT 는 Wikipedia 와 BooksCorpus 에서 레이블 없이 대량의 텍스트로만 사전학습 되었다
- 사전학습(Pre-train)
- 미세조정(Fine-tuning): 특정 과제에 특화된 모델을 만들기 위해, 파라미터를 추가하고 과제 데이터를 통해 학습하는 방법이다



3. 텍스트 정보를 반영한 딥러닝 기반의 추천 시스템

- 사용자 ID 와 강의 ID,
- 텍스트 데이터를 이용하여 강의를 추천해주는 모델
- 제안하는 모델은 예측모델(Prediction model)과 추천 모델(Recommender model) 두 가지



3. 텍스트 정보를 반영한 딥러닝 기반의 추천시스템

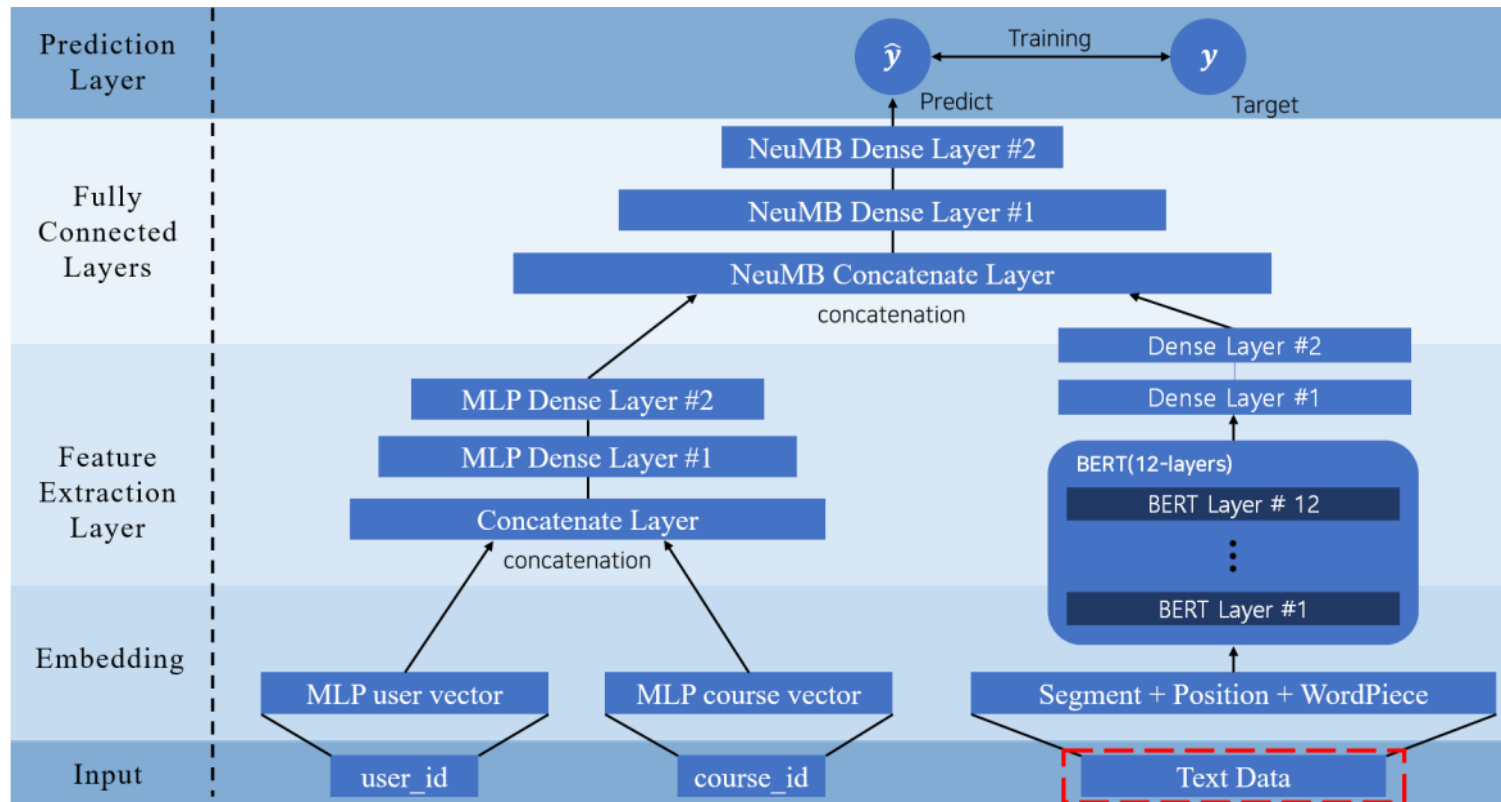
- BERT기반의 점수 예측 모델(BERT_PM)
- 자연어 처리 딥러닝 모형인 BERT 에 NCF 모델을 쌓은 형태를 이루는 점수 예측 모델 (BERT_based Prediction Model, BERT_PM)을 제안
- BERT_PM 은 사용자-강의 사이의 유의미한 특성을 추출하고, 텍스트 데이터로부터 추출된 특성과 함께 고차원적 관계를 포착하여 사용자가 강의에 매길 평가 점수를 예측한다
- 세 가지 항목의 데이터(사용자 ID, 강의 ID, 텍스트 데이터)를 입력값으로 받으며, 사용자 ID,강의 ID 는 원핫인코딩을 통해 입력된다.

두 항목은 희소한 벡터로 이루어져 있기에 밀집한 벡터로 변환하기 위해 임베딩 층(Embedding Layer)을 거쳐 잠재요인 벡터들로 표현된다.

잠재요인 벡터가 두 개의 Dense Layer 사용자-강의 사이의 유의미한 특성값이 추출된다. 텍스트 데이터는 사용자-강의 사이의점수 예측에 맞도록 Fine-Tuning 된 BERT 와 두 개의 Dense Layer 를 통해 텍스트 데이터로부터 특성값이 추출된다

3. 텍스트 정보를 반영한 딥러닝 기반의 추천시스템

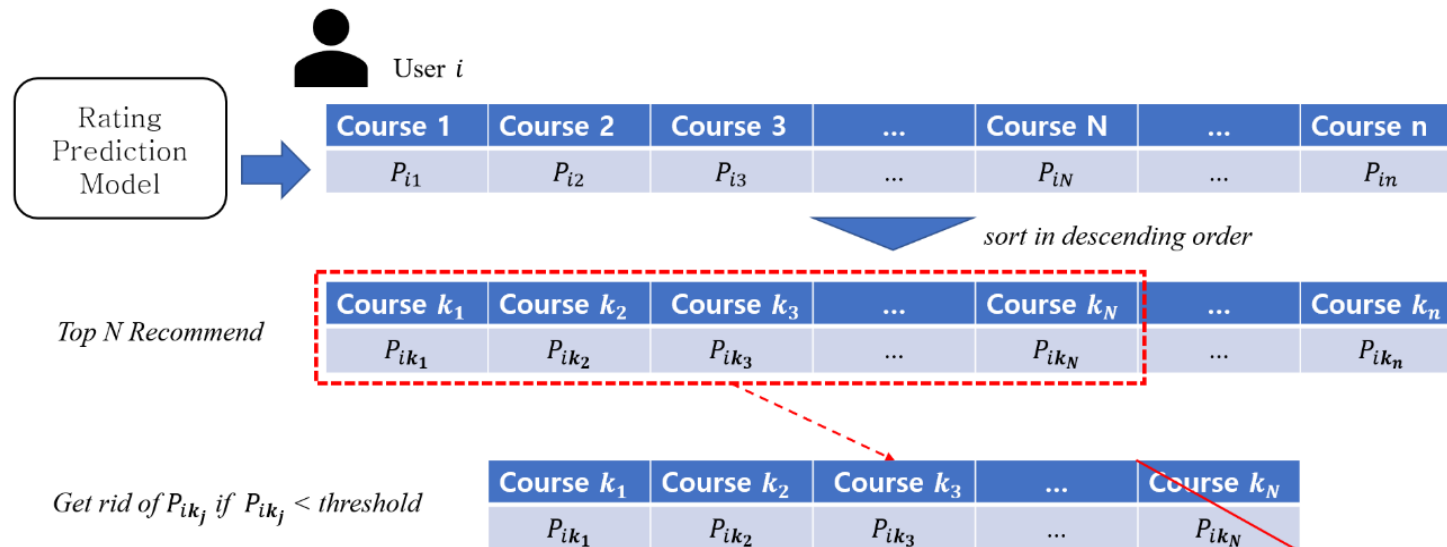
- BERT기반의 점수 예측 모델(BERT_PM)



3. 텍스트 정보를 반영한 딥러닝 기반의 추천시스템

추천 모델(Recommender model)

- 예측 모델로부터 계산된 예측 평가 점수를 이용
- 사용자가 들었던 강의를 제외하고 예측 평가 점수가 높은 순의 TopN 개의 강의를 추천한다
- 강의에 대한 예측 평가 점수가 일정 수준의 점수를 넘지 않으면 추천하지 않도록 함



4. 데이터 분석 및 실험

데이터 수집 및 전처리

- MOOC 온라인 플랫폼의 선두주자인 Coursera 에서 강의에 대한 정보와 사용자의 리뷰 및 평점 데이터를 이용하여 실험을 진행하였다 Coursera 에 존재하는 전체 강의 수는 7,279 개로, 원활한 실험을 진행하기 위해 과목별 필터링을 진행
- 11 개의 과목 중 IT 관련 3 가지 과목(ComputerScience, Data Science, Information Technology)을 선정하여 데이터를 크롤링을 했음
- 총 925 건의 강의로부터 331,240 건의 리뷰 데이터를 수집
- 수집된 데이터는 크게 사용자 리뷰와 강의 정보 데이터 두 가지고, 총 5 가지 세부 항목(사용자 ID, 강의 ID, 평점,리뷰 내용, 강의 설명)을 이용하여 데이터를 구성

No	User_id	Course_id	Rating	Review	Course_info
1	3243	203	5	The Course is Awesome...	Counting is one of the ..
2	1876	424	3	The content was indepth..	This course introduces...

4. 데이터 분석 및 실험

데이터 수집 및 전처리

사용자 리뷰를 모두 합친 데이터

강의의 평가 점수를 매기는 데 있어서 텍스트 데이터의 어떤 단어가 큰 영향을 주는지 확인할 수 있다.

3243

would be better off to use quizzes BUT I appreciate the use of puzzles as a complement to this course The instructors are enthusiastic and knowledgeable More emphasis can be given on simpler ways to put forward the ideas without tough use of mathematical terms All in all this was a good course. Thanks I have learned a lot .
The Course is awesome if you know some basic maths The instructors have a good way of teaching The problems in middle of videos help in understanding better The problems in tests are usually helpful in getting a better understanding.

4. 데이터 분석 및 실험

실험 프로세스

- 사용자 리뷰 데이터를 이용한 BERT_R(BERT-based Recommender System with User Review)
- 강의 설명 데이터를 이용한 BERT_C(BERT-based Recommender System with Course Description),

Recommender System	CF_user, CF_item	NCF	BERT_R	BERT_C
	↑	↑	↑	↑
Recommender model	Top N Recommender model with threshold 4			
Prediction model	Collaborative Filltering	NCF	BERT_PM	BERT_PM
Data sets (Details)	ratings_over5	ratings_over5, ratings_over3	ratings_over5, ratings_over3	ratings_over5, ratings_over3
	User_id, Course_id, Rating	User_id, Course_id, Rating	User_id, Course_id, Rating, Review_concat	User_id, Course_id, Rating, Course_info

4. 데이터 분석 및 실험

실험 방법 및 모델 학습

사용자가 자신의 선호도를 직접 표현하더라도 부정적 강의에 대해 끝까지 수강하거나 부정적인 피드백 리뷰를 남기는 행위는 긍정적 피드백에 비해 드물게 일어난다.

- 전체 데이터의 평점(Rating) 분포를 보면 대략 90%의 평점데이터가 4 점과 5 점에 편향되어 있고 평점의 평균은 약 4.5 정도로 계산된다

Data Set	Number of unique			Rating					sparsity
	User	Course	Review	1	2	3	4	5	
ratings over5	11,577	894	111,196	2,750	2,767	6,301	18,696	80,682	91.1(%)
ratings over3	27,691	910	164,988	3,863	3,861	8,944	27,351	120,969	94.2(%)
Raw data	165,814	925	331,240	7,810	6,946	16,713	54,894	244,877	98.0(%)

4. 데이터 분석 및 실험

실험 방법 및 모델 학습

- 예측 모델의 성능을 측정하기 위해 평가 측도로 MAE(Mean Absolute Error)와 RMSE(Root Mean Square Error)를 측정
- 추천의 정확성과 유효성은 추천시스템을 평가하기 위해 일반적으로 이용되는 지표인 F1@K 를 사용
- F1@K 는 Precision과 Recall 의 조화평균으로 두 지표를 동시에 반영한 지표로. Precision 은 모델이 사용자에게 추천으로 예측한 항목 중 실제 추천항목인 것의 비율이며, Recall 은 실제 사용자의 추천항목 중에 모델을 통해 추천된 항목의 비율
- 추천시스템이 얼마나 다양한 강의를 추천해 주는지 확인하기 위해 ItemCoverage@K 를 측정 Item Coverage 는 올바르게 추천된 고유한 아이템 항목의 수

$$\text{Precision@K} = \frac{\# \text{ of recommended items @k that are relevant}}{\# \text{ of recommended items @k}}$$

Precision@k : 추천한 강의 중에 사용자가 선호하는 강의를 얼마인가

Recall : 관련있는 강의(전체) 중 추천된 강의를 몇개인가

$$\text{Recall@K} = \frac{\# \text{ of recommended items @k that are relevant}}{\text{total \# of relevant items}}$$

$$F1@K = \frac{2(\text{Precision@K} \times \text{Recall@K})}{\text{Precision@K} + \text{Recall@K}}$$

4. 데이터 분석 및 실험

- 실험결과

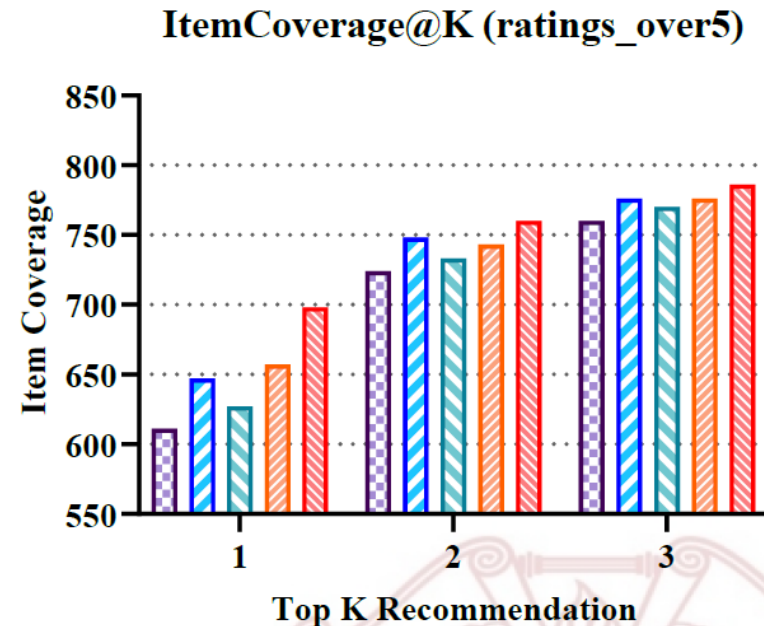
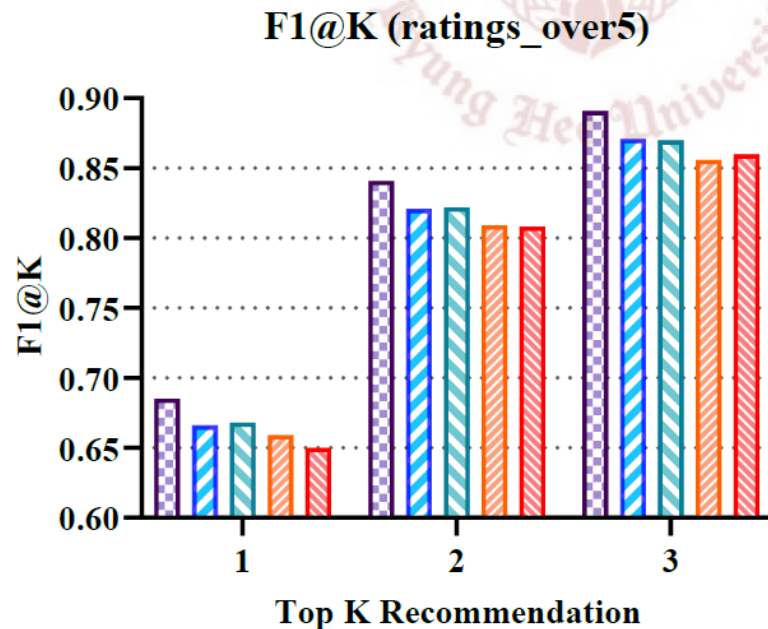
- 예측 모델 실험결과

- 제안한 BERT_R 모델은 MAE 값이 0.407 로 NCF 대비 약 2.6% 정도 감소

	ratings_over5					ratings_over3		
	NCF	BERT_R	BERT_C	CF_user	CF_item	NCF	BERT_R	BERT_C
MAE	0.454	0.435	0.436	0.586	0.554	0.418	0.407	0.419
RMSE	0.820	0.808	0.825	0.872	0.901	0.800	0.776	0.797

4. 데이터 분석 및 실험

- 실험결과
- 강의 추천 시스템 실험결과
- $NCF > BERT_R > BERT_C > CF_user > CF_item$ 순으로 F1 값이 높게 나옴
- BERT_R 모델의 경우 NCF 모델에 비해 Item Coverage 값이 16 개 높은 값을 보이면서 추천의 성능과
- 다양성을 모두 겸유함을 알 수 있음



4. 데이터 분석 및 실험

- 실험결과

- NCF 와 BERT기반의 예측 모델(BERT_PM)에 대해 예측 점수 도출까지 걸리는 시간을 확인해 보았을 때, NCF 는 약 3.04 초가 소요 BERT_PM 은 약 162.70 초가 소요.
- BERT 기반의 점수 예측 모델은 12 개 층을 가지는 BERT 를 통해 텍스트로부터 고유 특성을 추출해야 하기에 NCF 의 모델에 비해 소요시간이 오래걸림
- 사용자의 정보가 적을수록 수강 강의가 겹치기 어려울 것 그러나 사용자 리뷰는 다른 강의일 경우라도 강의의 특성에 대한 사용자의 표현은 겹치는 것이 가능하며, 더 많은 유사 사용자 탐색을 가능하게 해 준다. 즉, 이는 사용자에게 대한 정보가 적을 때 텍스트 정보를 반영하는 것이 더 좋은 성능의 추천을 가져 온다는 것을 의미함

5. 결론

- 정보가 적은 사용자들이 추가되었을 경우 본 연구에서 제안한 모델의 성능이 가장 높게 나왔으며, 사용자 리뷰를 이용한 BERT_R 모델에서 협업 필터링과 NCF 기반의 추천시스템보다 다양한 추천이 성공적으로 이루어지는 것을 확인하였음
- 다만 예측 점수를 도출하는 시간은 BERT 기반의 모델이 NCF 모델에 비해 오래 걸린다는 단점이 존재함
- 리뷰라는 시스템이 정보가 적은 사용자들이 추가 되었을 때 모델의 성능이 높게 나온다는 점에서 rating뿐만이 아닌 리뷰의 추가 정보로 사용자들을 파악할 수 있는 점이 흥미로웠고 추천 시스템이 개발 됨에 있어서 기업들의 리뷰 작려 ex) 네이버 페이 가 단순 별점-소비자 선택이 아니라 상품-소비자로 이어져 추천시스템-기업에서 리뷰의 의의를 확인 할 수 있을 것 같음