1. 서론

1.1 연구 배경

정보 통신 기술의 발달로 수많은 제품과 콘텐츠들이 생산된다. 이에 결정권자인 소비자를 돕는 추천 서비스에 대한 관심이 점차 대두되고 있고 이미 여러 분야에서 광범위하게 활용되고 있다. 동영상 플랫폼인 넷플릭스(Netflix)의 경우, 가입 시 관심 있는 장르를 선택하게 한다. 이후 시청한 영상 목록과 영상에 대한 평가를 반영하여 시청자 개인의 취향에 맞는 영상을 추천하는 방식으로 서비스를 제공하고 있다. 이외에도 뉴스 기사, 온라인 쇼핑 등 다양한 분야에서 추천 서비스에 대한 적용이 활발하게 이루어지고 있다.

최근 개인 맞춤형 추천 서비스를 대학 도서관에 도입하려는 움직임이 활발하다. 2019년 서울대학교 중앙도서관은 'S-Curation'이라는 명칭의 개인화 서비스를 제공하기 시작했다. 해당 시스템은 학생의 도서 대출 이력과 본인이 등록한 관심 키워드를 바탕으로 도서 취향을 파악한다. 이를 바탕으로 개인별로 맞춤형 도서 및 보고서를 추천해주는 방식으로 운영된다. 또한 이화여자대학교에서는 2017년부터 제공해 온 '독서 프로파일링 서비스'를 2020년 개편했다. 추천 도서의 범위를 전자책까지 확대하고, 추천 도서가 특정 분야로 편중되지 않게 하기 위하여 관심 분야를 관리하는 기능도 추가했다.

2020년 11월, 성균관대학교 학술정보관에서 독서진흥 프로젝트를 담당하시는 이규성 사서님과, 북큐레이션을 담당하시는 정진아 사서님을 한 차례 인터뷰한 바 있다. 그를 통해, 우리 대학 역시 개인화 추천 서비스 시행에 관심이 있음을 알 수 있었다. 현재 학술정보관은 도서를 검색했을 때 해당 도서의 주제를 기반으로 하여 '동일주제 인기자료'를 추천하는 정도에 머물러 있는 상황이다. 이외에도 현재 시행중인 북큐레이션은 수작업으로 도서를 선정 및 전시하는 방식으로 서비스가 제공되고 있다.

이에 따라, 본 연구는 성균관대학교 학술정보관에 적용 가능한 개인별 도서 추천 서비스를 구현하는 것을 목표로 한다. 더 나아가, 개인의 대출 기록을 바탕으로 취향에 맞는 도서를 추천하는 새로운 서비스의 도입은 성균관대학교 학우들의 독서 진흥에 기여할 것으로 기대한다.

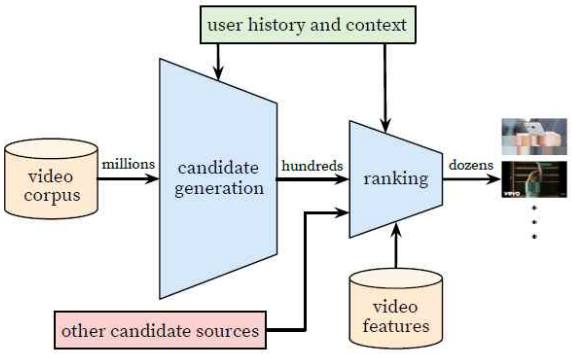
1.2 연구 동향

추천시스템은 다양한 분야와 기법을 통해 구현되고 있으며, 크게 협업 필터링(Collaborative Filtering), 행렬 분해(Matrix Factorization), 그리고 딥러닝(Deep Learning)을 사용한 방식으로 분류할 수 있다.

조연선(2019)은 협업 필터링 방식을 사용해 도서 추천 시스템을 구축했다. 특정 고등학교 재학생들의 대출 기간을 이용하여 평균 일수을 기준으로 선호도 점수를 1점부터 5점 사이의 값으로 부여했다. 즉, 평균 대출 일수보다 대출일수가 길다면, 높은 선호도를 부여하는 방식이다. 유사도를 구하고 이웃한 사용자를 생성해내는 협업 필터링 방식을 통해 예측 선호도 순위대로 10권을 추천하도록 설정했으며 도서 목록에 대한 만족도를 매우 만족(5), 만족(4), 보통(3), 불만족(2), 매우 불만족(1)의 5점 척도로 평가했다.

정승윤(2017)은 기업의 데이터를 대상으로 추천 시스템의 최적화에 대한 실증 연구를 진행했다. 추천 시스템 평가를 위한 알고리즘은 사용자와 아이템 간의 상호작용 특징 벡터를 잘 표현하는 Factorization Machine을 사용했다. 추천 시스템 최적화를 위해 SGD(Stochastic Gradient Descent) 방법을 이용하여 최적의 파라미터를 구했다. 또한, 원래 데이터 행렬에서 값이 없었던 부분을 제외한 후 에러를 계산하고 최소화시키는 것, 즉 RMSE 값을 줄이는 것이 추천 문제라고 할 수 있음을 주장하였으며, 실제 기업 데이터 평점 추천 모델에 적용, 최적의 FM을 설정했다.

딥러닝 모델을 활용한 추천 시스템의 연구에는 유튜브의 영상 추천시스템이 있다. 이 연구는 유튜브의 방대한 이용자(user) 정보와 영상(video) 정보를 이용해 깊은 신경망(deep neural network) 기반의 영상 추천 시스템을 개발했다. 두 가지 신경망으로 모델을 구성해, 각 단계의 네트워크에서 유튜브의 대량의 영상 범위를 줄이며 이용자 개개인에게 적은 개수의 우수한(rich) 영상 리스트를 제공한다. 첫 번째 신경망은 후보자 생성(candidate generation)으로, 방대한 유튜브 데이터를 이용해 몇 백 개의 영상만을 이용자 별로 제한한다. 두 번째 신경망은 순위 네트워크(ranking network)로, 이용자와 영상 정보를 통해 이전 신경망을 통해 제공된 영상 리스트에 선호도 점수를 부여해 최종적으로 이용자에게 상위 점수의 영상을 추천한다. 방대한 영상 뭉치(video corpus)에서 이용자에게 영상을 추천하기 위해, 이용자의 영상 시청 기록(watch history)과 검색 기록(search history)을 임베딩 값으로 표현한다. 영상 시청 기록과 검색 기록 임베딩은 이용자의 기본적인 정보 등과 어우러져 이용자를 표현해 영상 추천의 입력값(input)으로 사용된다. <그림 1>을



〈그림 1〉 유튜브 영상 추천 시스템 아키텍처

통해, 유튜브 추천시스템의 모델 아키텍처(architecture)를 확인할 수 있다(Paul, Jay, & Emre, 2016).

2. 연구 방법 및 전처리

2.1 연구 방법

2.1.1 행렬 분해

행렬 분해(Matrix Factorization)란 User-Item Matrix를 User Latent Matrix와 Item Latent Matrix의 행렬 곱으로 분해하는 행렬 인수분해 방법이다. 사용자가 평가하지 않은 아이템에 대한 선호도를 쉽게 추정할 수 있는 방법으로 SVD, SGD, ALS, 마르코프 체인과 같은 알고리즘들이 주로 사용된다.

SVD(Singular Value Decomposition)는 특이값 분해 방법이다. 이는 행렬의 전체적인 구조를 기반으로 추천을 하는 방식이다. 차원 축소 개념의 일종으로 고차원 행렬을 저차원의 행렬로 축소시켜 분석의 정확성을 높이고 계산 속도를 향상 시킬 수 있다. SVD는 기존 행렬을 직각 행렬 2개와 1개의 대각 행렬로 분해한다. 추천 시스템에서는 각각의 분해된 행렬을 이용하여 차원 축소를 실시할 수 있다. SVD는 선호도 정보가 있을 경우 이를 바탕으로 행렬을 분해하였을 때, 잠재요인(latent factor)를 잘 정의한다고 알려져 있다(정승윤, 2017).

SGD(Stochastic Gradient Descent) 방법은 실제 선호도와 예측된 선호도의 차이를 에러로 정의하며 결측치와 상관없이 선호도 계산을 진행한다는 장점이 있다. 우선, 값이 존재하는 선호도를 이용해 순차적으로 Gradient Descent 과정을 진행하며 오차를 계산한다. 이 과정에서 편미분을 통해 User Latent과 Item Latent 의 업데이트를 진행해 선호도를 산출한다.

2.1.2 사용자 기반 협업 필터링

사용자 기반 협업 필터링(User-based Collaborative Filtering) 알고리즘은 사용자의 아이템에 대한 선호도를 기반으로 유사한 사용자들의 그룹을 찾아낸다. 유사한 사용자 그룹에서 선호도가 높은 아이템들을 추천해주는 방식이 사용자 기반 협업 필터링 알고리즘을 사용한 추천 방식이다. 사용자 기반 협업 필터링에서 가장 중요한 부분은 사용자 사이의 유사도를 측정하는 것이다. 사용자가 아이템에 대하여 남긴 선호도 기록을 활용하여 사용자와 사용자 간의 유사도를 계산하고 이를 기반으로 알고리즘을 구현한다. 유사도 측정에는 다양한 방법들이 있다. 피어슨 상관관계 기반 유사도 측정, 유클리드 거리 기반 유사도 측정, 코사인 유사도 측정 등이 있다. 유사도 측정 방법에 따라 추천 시스템이 성능이 달라질 수 있으므로, 추천 환경에 맞는 유사도 방법을

적절하게 사용하여야 한다(조연선, 2019).

2.1.3 신경망(Neural Network)

신경망은 하나 이상의 계층으로 구성된다. 그리고 각각의 계층은 여러 개의 노드를 포함한다. 이 때, 각 계층의 노드들은 하위 계층 노드들의 출력값들을 가중 합(weighted sum)의 형태로 결합한다. 하위 계층일수록 낮은 수준들을 추출하고, 상위 계층으로 갈수록 높은 수준의 특징을 추출하여 결합하는 특성에 의해, 신경망을 구성하는 계층의 수가 많을수록 더욱 정교한 수준의 특징을 추출할 수 있게 되고 복잡한 작업을 수행하는데 있어 탁월한 성능을 보이게 되는 것이다(김인중, 2014). 여러 층이 쌓인 깊은 신경망(Deep Neural Network)의 경우, 각 계층에서 훈련 데이터의 특징을 추출하고 더 높은 수준으로 이를 추상화하는 과정을 거치며 별도의 특징 추출 알고리즘 없이 신경망이 스스로 패턴을 인식할 수 있다.

2.1.3.1 다중 입력 모델

2.1.3 절에서 언급한 바와 같이 신경망은 하나의 모델이 데이터를 학습하여 결과값을 출력하기 때문에 데이터를 모델에 맞게 구성하여 입력하면 된다. 하지만 모든 데이터가 하나의 동일한 모델에 적합하게 구성되지 않는 경우가 존재한다. 따라서 데이터의 특성에 알맞은 모델을 적용하여 학습시킨 후, 이를 모두 결합하여 최종 모델을 구성하는 방식을 채택해야 한다. 이러한 경우를, 다중 입력 모델(Multi Input Model)이라고 한다.

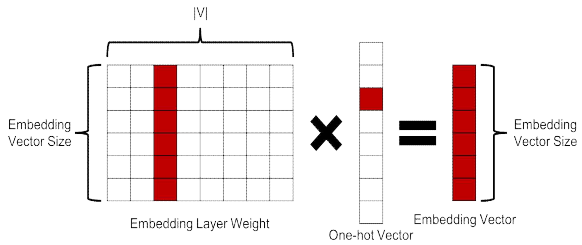
네이버 쇼핑의 상품 카테고리 자동 분류의 경우가 이에 해당한다. 카테고리 자동 매칭을 위해 판매사로부터 상품명과 상품의 이미지를 비롯한 다양한 정보를 제공받아, 상품명(자연어)을 사용한 모델과 이미지를 사용한 모델을 별개로 구축해 학습한다. 이에 그치지 않고 두 모델을 하나로 결합하여 딥러닝 모델을 학습시킨다(네이버 쇼핑 플랫폼, 2019).

2.1.4 임베딩(Embedding)

데이터의 특성에 따라 전처리를 완료하면 다음으로 모델에 넣어 학습을 진행하게 된다. 이때 수치형 데이터가 아닌 경우, 그대로 모델에 넣을 수 없다. 그래서 데이터의 특징을 추출하여 수치화를 벡터화가 필수불가결하다. 이러한 벡터화 과정을 임베딩이라고 한다.

임베딩의 대표적인 방법으로는 원핫인코딩(one-hot-encoding)이 있다. 남자와 여자, 미성년자와 성인처럼 카테고리로 분류되는 변수의 경우 원핫인코딩을 통해 수치화한다. 그러나 이 방식은 카테고리의 수가 많아지면 벡터 공간이 매우 커지기 때문에 비효율적이다. 또한 어떤 카테고리에 속하는지는 표현할 수 있지만 특징을 표현해주지

못한다.



〈그림 2〉 임베딩 계층의 동작 원리

그래서 자연어와 같이 카테고리가 많아지는 경우에는 위와 같은 밀도가 낮은(sparse) 기법보다는 데이터들의 밀도를 높이도록 원핫벡터와 가중치의 내적을 구하여 차원이 낮은 벡터 공간에 사영(projection)시켜주는 방법을 사용한다(김기현, 2019). Word2Vec이 그 대표적인 예이다. 이렇게 얻은 임베딩 행렬 자체를 신경망에 입력하여 학습시키면 가중치들을 학습하여 임베딩 행렬을 함께 업그레이드 하고 결과적으로 비슷한 특성들은 벡터 공간에서 가까운 위치에 놓이게 되는 것이다(Yong, T., Hazarike, D., Poria, S., Cambria, E., 2018).

2.1.4.1 ROBERTa

자연어의 경우, ROBERTa의 사전 학습(pretrain)된 모델을 사용해 임베딩을 표현했다. ROBERTa(A Robustly Optimized BERT Pretraining Approach)는 2019년에 발표된 모델로, 2018년 Google에서 발표한 트랜스포머(transformer) 기반의 사전 훈련(pretrained)된 모델인 BERT(Bidirectional Encoder Representations from Transformers)의 변형된 형태이다. ROBERTa는 BERT를 개선하기 위한 목적으로 다수의 실험을 통해 생성된 모델로, 일반적인 BERT의 방식과 달리 next sentence prediction을 생략하고 Masked Language Model(MLM)만을 이용, 정적 마스킹(static masking) 대신 동적 마스킹(dynamic masking)을 사용하고 배치(batch) 사이즈를 더 키우는 등 BERT에 변화를 주어 성능을 높인 모델이다(Yinhan, Myle Ott, Naman Goyal, Jingfei, Mandar Joshi, Danqi, Omer , Mike , Luke, & Veselin, 2019).

2.1.4.2 특성 추출과 EfficientNet

특성 추출(Feature Extraction)은 원시 데이터(raw data)로 부터 유의한 정보들만을 이용해 기존 차원보다 낮은 차원으로 축소하는 것을 의미한다. 이는 데이터를 차원이 낮은 벡터 공간에 사영하는 임베딩과 유사하며, 이미지 임베딩을 이미지 특징 추출이 대체할 수 있다고 해석해도 무방하다. 이미지의 특성을 추출할 수 있는 기법은 다양하지만, 최근 많이 사용하고 있는 기법은 ConvNet (Convo- lutional Neural Network)을 이용하는 방법이다. 본 연구에서는 효과적인 추출을 위해 ConvNet 기반의 EfficientNet을 전이 학습하여 이미지 특성 추출을 진행했다. EfficientNet은 기존의 ConvNet을 발전시키기 위해 시도되었던 depth scaling, width scaling, resolution

scaling을 동시에 고려해, 세 가지 요소의 균형을 맞추어 scaling up을 하는 compound scaling을 방식을 적용, 이미지 분류에서 좋은 성능을 보이고 있다. (Tan & Quoc, 2019).

2.2 데이터 전처리

2.2.1 공통 데이터 전처리

본 연구에서는 성균관대학교 학술 정보관의 2015년부터 2019년도의 학부생 대출 기록 데이터를 사용했다. 대출 기록은 대출자의 정보, 대출 도서의 정보, 대출이 이루어진 시점과 장소에 대한 정보로 이루어져있었다. 그 중 대출자 번호, 대출자 소속, 도서명, 청구기호, 대출일자, 반납일자만을 예측에 사용했다. 또한, 자격증 시험 대비 문제집과 같은 취업에 도움이 되는 서적들과 과제도서로 분류된 도서들이 개인의 기호를 반영한다고 보기에는 무리가 있어, 해당 도서들은 제외했다. 또한, 한 학생의 도서 선호도 예측을 위해서는 충분한 도서 대출 기록이 있어야 하므로, 도서를 1권 대출한 학생과 소속 학과가 없는 학생들의 기록은 삭제했다.

대출자 번호의 경우, 개인정보의 유출을 방지하기 위해 난수 처리된 상태로 제공받았으며, 단과 대학, 특정 학과 등의 다양한 기준으로 분류되어 있던 대출자 소속은 단과대 단위(문과대학, 사회과학대학, 경영대학 등)로 재분류했다.

도서관 내의 도서는 이용자들이 동시에 동일한 도서를 대출할 수 있도록 복본을 비치해 둔다. 복본은 서지정보, 판 사항, 제목 등 모두 동일하지만 청구 기호를 통해 복본인 것을 나타낸다. 모든 도서는 고유한 등록번호를 갖고 있어 본 연구에서 등록번호를 통해 도서를 구별하는데, 복본을 동일한 도서로 인식할 수 있도록 기준이 되는 도서

〈표 1〉 성균관대학교 학술정보관 제공 데이터의 변수명과 내용

변수명	내용	변수명	내용
등록번호	학술정보관 도서 등록번호	대출처리일자	대출이 처리된 날짜
학번난수	난수처리한 학번	반납처리일자	반납이 처리된 날짜
소장분관	도서가 소장된 분관	대출상태	대출중 / 반납
소장서고	도서가 소장된 서고	대출유형	장기대출 / 대출취소
서명	도서 제목	반납유형	정상반납 / 연체반납
청구기호	도서 청구기호(DDC)	서지번호	도서 정보를 식별하는 번호
전공	대출자의 소속 학과		

와 복본의 등록번호를 동일하게 매칭시켜 주었다.

2.2.2 도서 선호도 계산

성균관대학교 학술정보관에는 선호도를 파악할 수 있는 지표가 없기 때문에 선호도를 계산하기 위한 식을 직접 도출해야 한다. 본 연구에서는 책 장르, 대출 기간을 이용하여 선호도를 계산하는 방법을 제안한다.

2.2.2.1 도서 장르에 의한 선호도

학생들이 책을 고를 때 보통 관심 있는 주제의 책을 선정하기 때문에 가장 핵심적인 선호도 기준이 될 수 있다고 생각한다. 책 장르는 학술정보관의 경우 듀이 십진 분류표(DDC)를 이용하기 때문에 DDC를 기준으로 학생의 장르 선호도를 파악할 것을 제안한다. DDC 기준으로 10의 자리 숫자까지 같은 책들을 같은 장르의 책으로 판단하기로 했다. 예를 들어 8로 시작하는 책은 문학으로 분류된 책이다. 두 번째 자리가 2인 경우 영미문학으로 82X인 경우 영미 문학에 속하는 도서이다. 같은 문학 책이더라도 세부적인 취향 차이가 있을 수 있기 때문에 10의 자리 수까지 같은 책을 동일한 장르로 판단했다. 해당 대출자가 대출한 전체 도서의 장르 중에서 해당 도서의 장르가 몇 퍼센트를 차지하는지를 파악하여 이를 5개의 구간으로 나누어 선호도 점수를 매기는 방법을 제안한다.

주류표 (first summary : ten main classes)	강목표 (second summary : hundred divisions)
800 문학 (Literature)	800 문학 (Literature, rhetoric & criticism)
	810 동양문학 (Oriental literature)
	820 영미문학 (English literature)
	830 독일문학 (German literature)
	840 프랑스문학 (French literature)
	850 이탈리아문학 (Italian literature)
	860 스페인문학 (Spain literature)
	870 라틴문학 (Latin literature)
	880 고전희랍문학 (Classical & Greek literature)
	890 기타 제문학 (Other literature)

〈그림 3〉 DDC 주류표와 강목표 예시

2.2.2.2 대출 기간에 의한 선호도

성균관대학교 학술정보관 학부생들의 대출 가능 기간은 14일이며, 학부생들의 대출 기간의 평균은 14일이었다. 따라서 개인의 도서 이용 행태에 따라 대출 기간이 달라지므로 단순히 대출 기간만으로 선호도를 판단하기는 어렵다고 생각하여 개인의 평균 대출 기간을 이용한 식을 도출했다.

$$\text{대출기간 선호도} = \frac{\text{특정도서 대출 일수}}{\text{해당 대출자의 평균 대출 일수}}$$

우선 특정 대출자가 평균적으로 책을 읽는 평균 대출 일수를 구한다. 그 다음 특정 도서를 대출자가 몇 일 동안 읽었는지를 구한다. 만약 특정 도서를 대출자의 평균 대출 일수만큼 대출을 했다면 선호도는 3, 평균 이하로 했다면 구간에 따라 1~2, 평균 이상으로 했다면 4~5점을 부여한다.

2.2.3 딥러닝 데이터 전처리

딥러닝을 이용한 도서 개인화 추천 모델은 대출자가 대출한 도서 목록과 대출자의 소속을 이용해 가장 마지막에 대출한 도서를 예측했다. 도서의 정보와 대출자의 정보가 충분해야 예측이 잘 이루어질 수 있는데, 대출자 관련 변수의 경우 개인정보 유출의 위험으로 변수 수집에 한계가 있어 도서의 정보를 추가적으로 수집했다. 성균관대학교 학술정보관 홈페이지에서 각 도서 검색결과를 웹 스크레이핑(Web Scraping)하여 출판연도와 쪽수, 책 표지를 학습 데이터에 추가했다. 이 과정에서 도서 표지, 쪽수, 출판 연도가 적절하게 스크레이핑 되지 않은 도서는 학습 데이터에서 제외했다.

도서의 청구기호는 도서의 주제와 내용을 기호화한 DDC로 이루어져있는데, DDC의 첫번째 숫자는 도서를 분류하는 가장 큰 대분류를 나타내어, 청구기호의 첫 숫자가 0이면 '총류', 1이면 '철학', … 9이면 '역사'의 주제를 갖고 있다고 해석할 수 있다. 이를 이용해, 도서의 주제를 0부터 9의 범주형 변수로 표현했다.

우선적으로 Label Encoder를 이용해 등록번호 Labeling을 진행했다. 수치형 데이터로 변형하기 위해 도서 제목 등의 자연어 변수와 이미지, 범주형 변수는 전이학습을 통해 임베딩 레이어를 산출했다. 도서 페이지수와 출판연도 등의 수치형 데이터는 제곱근과 제곱을 취하는 등의 변형을 가해 변수를 추가했다. 이후 MinMaxScaler를 통해 정규화(Normalization)를 진행하여 데이터의 범위를 일치시켰다.

본 연구에서는 대출 기록 데이터와 대출자 정보 데이터, 2개의 데이터를 사용하는 다중 입력 모델을 사용한다. 딥러닝 모델의 예측 목표가 각 대출자의 대출 기록을 통해 가장 마지막에 읽은 도서를 예측하는 것이므로, 대출 기록 테이블의 경우에는 대출자 번호를 기준으로 대출 기록을 목록화 했다. 대출자 테이블은 대출자 번호와 해당 대출자가 속한 단과대학 총 2개의 변수로 이루어져있다.

3. 연구 설계

3.1 선호도 기반 모델

3.1.1 행렬 분해

2.1.1 절에서 설명한 SGD와 SVD 기법을 사용해 모델링을 진행했다. SVD 모델을 사용한 결과 User-Item Matrix와 같이 결측값이 많은 희소 행렬에서는 좋은 성능을 보이지 않았다. SVD 모델은 단순한 형태이기 때문에 학부생들이 많이 읽은 810번 대의 문학 장르 베스트셀러들이 추천 도서 목록으로 자주 도출되어 유의미한 결과를 얻지 못했다.

SGD 알고리즘을 사용한 결과 RMSE는 0.0032 이 도출되었다. SVD와 다르게 오차

를 계속해서 업데이트하여 반영하기 때문에 베스트셀러만 추천 목록으로 도출되는 문제점을 해결할 수 있었다. SVD 모델과 달리 대출자의 장르 선호도를 반영했다는 점에서 유의미한 결과가 도출됐다.

3.1.2 사용자 기반 협업 필터링

최종 선호도 설정 후, 대출자와 대출자 사이의 유사도 계산을 위하여 학생들의 책에 대한 선호도 행렬을 만들었다.

최종 선호도 행렬에 대하여 코사인 유사도 측정 방식을 사용하여 대출자 사이의 유사도 행렬을 만들었다. 만들어진 유사도 행렬에 기반하여 특정 대출자의 번호를 입력하여 이와 유사도가 높은 대출자 10명이 포함된 그룹을 추출해내었다. 해당 그룹이 읽었던 도서 10개를 추천해주었다. 추천 순서는 대출자의 유사도 높은 순서로 그룹 내 대출기록이 있는 도서 10권을 추천해주었다. 추천 도서에는 당사자가 읽은 도서를 제외했다.

3.2 딥러닝 모델

3.2.1 모델 아키텍처

본 연구의 딥러닝 모델의 경우, [별첨 1]의 아키텍처로 기본 딥러닝 모델을 구성했다. 2.1절에서의 도서 대출 데이터(book table)와 대출자에 대한 정보를 포함하는 데이터(user table)를 입력 값으로 사용한다. 이때, 도서의 정보와 대출자의 정보를 일괄 포함하기 위해 다중 입력 모델이 이용된다. 도서 대출 데이터의 입력 값은 도서의 제목, 이미지, 장르 그리고 그 외 도서의 페이지와 출판년도 정보를 포함한 임베딩 레이어의 참조 테이블(look up table)로서 사용된다. 대출자 데이터의 경우, 전공에 대한 임베딩 레이어의 참조 테이블로서 사용된다. 도서에 대한 다양한 정보 그리고 대출자에 대한 정보를 포함한 임베딩 레이어들을 각각 평균 풀링(average pooling)하여 하나로 병합했다(concatenate). 즉, 각 대출자와 개인이 대출한 도서 데이터(이미지, 자연어 등)가 하나의 정보로 표현된다. 완성된 데이터는 다양한 레이어를 거치며, 최종 dense 레이어에서 도서의 개수(133878 개)로 출력되어 분류한다. 이때 categorical cross entropy의 비용 함수(loss function)를 이용해 softmax로 다중 클래스(multi class)를 분류했다.

3.2.2 임베딩 레이어

도서의 정보를 다각도로 반영하기 위해, 네 가지 임베딩 레이어를 이용했다. 첫 번째 임베딩 레이어는 Bert book title embedding이며, RoBERTa의 문장 트랜스포머(sentence transformer)를 이용해 임베딩 값을 산출했다. 본 연구의 경우, 도서 제목이 한국어, 영어, 중국어 등 다양한 언어가

포함되어 다국어 임베딩이 가능한 XLM(Cross-lingual Language Model Pre-training)를 사용했다. 두 번째 임베딩 레이어는 efficient net book image embedding으로 EfficientNet의 사전 훈련된 모델을 이용하여 임베딩 값을 산출했다. 세 번째 임베딩 레이어는 book genre embedding으로 16차원의 장르를 5차원으로 축소하여 임베딩한 값을 이용했다. 네 번째 임베딩 레이어는 book information embedding으로 책의 정보: 출판 연도와 페이지 수를 이용해 임베딩 값을 생성했다. 학생의 정보를 반영하기 위해, 한 가지 임베딩 레이어를 이용했다. 이는 major embedding 으로 도서 제목의 임베딩 방식과 동일하게 진행했다. <그림 4>는 RoBERTa를 통해 산출된 도서의 제목 임베딩 값을 시각화(visualization)한 것이며, [별첨 2]는 코사인 유사도가 큰 순서대로 도서 제목을 나열한 것이다.

3.2.3 모델 학습

딥러닝 모델의 입력 값은 2.2.3절에서 전처리한 대출 기록 데이터와 학생 정보 데이터이다. 대출 기록 데이터의 경우, 대출자의 대출 목록을 포함하며 학생 정보 데이터는 대출자의 전공을 포함한다.

모델의 학습 단계에서는 정확도(accuracy)를 사용해 학습을 진행했으며, 정확도는 전체의 데이터 샘플(sample) 수를 올바르게 예측한 샘플의 수로 나눈 것이다.

〈그림 4〉 도서 제목 임베딩 시각화

[별첨 1]의 기본 모델 아키텍처를 기반으로 세 개의 변형 모델을 설계했다. 이는 과적합(overfitting)을 방지하기 위해 실행했으며, 기본 아키텍처에서 변형된 세 개의 딥러닝 모델을 [별첨 3]에서 확인할 수 있다. 기본 딥러닝 아키텍처와 달리, flatten layer, batch normalization layer 추가 또는 옵티마이저(optimizer) 변경 등의 변형된 모델을 생성했다. 정확도(accuracy)가 95% 이상인 세 개의 모델을 선정했으며, 투표(voting) 앙상블 기법의 soft vote을 이용해 최종 모델을 선정했다.

4. 연구 결과 및 평가

SVD, SGD, 사용자기반 협업필터링(UF) 그리고 딥러닝(DL)을 이용한 네 가지 모델

의 평가를 6명의 학생 대상으로 사용자 검증을 진행했다. 학생들의 성균관대학교 학술정보관 대출 이력을 입력값으로 넣어 각 모델에서 도출한 예측값 중 상위 5개의 도서를 추출했다. 추출한 도서 중 읽고 싶은 책에는 1점, 그렇지 않은 책에는 0점을 부여하는 방법으로 평가했다.

학생들의 평가를 바탕으로 각 모델의 사용자 만족도를 평균 내어 산출한 결과는 <표 2>와 같다. 학생들의 추천 목록에 대한 만족도가 가장 높았던 모델은 딥러닝 기반의 모델로 5점 만점에 4점을 기록했다. 기존에 베스트셀러를 중심으로 추천 목록을 도출해 대출자 개개인의 선호도를 반영할 수 없을 것이라고 예상했던 SVD 모델은 두 번째로 높은 만족도를 보였다. SVD에 비해 개개인의 선호도를 반영할 것으로 예상했던 SGD와 사용자기반 협업 필터링의 경우 3.16점으로 다른 모델에 비해 낮은 만족도를 보였다.

〈표 2〉 사용자 검증 결과

사용자	SVD	SGD	UFC	DL
학생 1	3	4	1	4
학생 2	3	4	4	5
학생 3	5	2	3	3
학생 4	5	2	4	5
학생 5	3	3	3	4
학생 6	4	4	4	3
총점평균	3.83	3.16	3.16	4

5. 결론

본 연구에서는 성균관대학교 학술정보관의 대출 기록 데이터를 이용하여 개인화 맞춤 도서 추천 시스템을 구현했다. SVD와 SGD 기법을 이용한 행렬 분해 모델, 사용자 간의 유사도 계산을 통해 도서를 추천하는 사용자 기반 협업필터링, 그리고 대출자의 대출 기록을 이용해 이후의 대출 도서를 예측하는 딥러닝 모델, 총 4가지 모델링을 진행했다. 객관적인 모델의 비교 및 평가를 위해 성균관대학교 학술정보관의 이용자를 대상으로 사용자 만족도를 조사하여 결과적으로 가장 만족도가 높은 모델은 5점 만점의 4점으로 딥러닝 기법을 사용한 것이었고, SGD와 사용자 기반 협업 필터링은 3.16점으로 가장 낮은 만족도를 보였다.

일반적으로 독서의 특성상 영화와 다르게 별점을 매기는 등의 사용자 선호도 측정 과정이 존재하지 않기 때문에 그러한 점을 극복하고, 대출 기록만을 이용해 다양한 방식으로 개인화 맞춤 도서 서비스를 구현하려고 했다는 점에서 의의가 있다. 그러나 딥러닝 모델의 경우, 최종 예측 대상이 도서관에 있는 모든 도서이다 보니 범위가 매우 넓어 과적합의 가능성이 매우 높다. 향후, 예측 대상을 더 작은 범주로 조정하거나 다른 기준(도서의 장르 등)을 적용하는 방식으로 모델의 타당성을 높일 수 있을 것으로 전망한다. 머신러닝 모델링의 경우, 모델 구축 환경의 미비로 2019년도 대출기록만을 사용했다는 점에서 한계가 있다. SGD

모델링의 경우 이러한 문제를 해결해줄 Parallel SGD 알고리즘인 DSGD, FPSGD 등을 활용해 볼 것을 이후의 개선 방안으로 제시한다. 협업 필터링 모델의 경우, 사용자 간의 유사도만을 측정했기 때문에 아이템의 유사도를 반영하지 못했다는 한계점이 존재한다. 추후 아이템 기반의 협업 필터링 모델과 결합하여 모델의 성능을 향상시킬 수 있을 것이다. 또한, 구현한 모델들이 실제적으로 효용성이 있는지에 대한 사용자 검증을 진행하였으나, 표본이 충분하지 않다는 한계가 있다. 추후 다양한 소속의 학술정보관 이용자들을 대상으로 추가적인 검증을 진행함과 동시에, 평가지표를 보완하여 모델의 효용성 평가를 개선할 수 있을 것이다.

다양한 분야와 플랫폼에서 사용되고 있는 추천시스템은 도서관계에서도 활발하게 연구 및 도입이 이루어지고 있다. 그러나 다른 분야와 달리 이용자가 직접적으로 평가하는 도서 선호도가 없다는 점에서 어려움을 겪고 있다. 본 연구는 사용자의 추가적인 선호도 입력 없이 대출 기록만을 이용해 도서 추천 시스템을 구현하였으며, 앞으로 더 효율적이고 효과적인 개인화 도서 맞춤 시스템의 발전을 위한 토대가 될 것으로 전망한다.

참 고 문 헌

김기현 (2019). 김기현의 자연어 처리 딥러닝 캠프. 서울: 한빛미디어

김인중 (2014). Deep Learning: 기계학습의 새로운 트랜드. 한국통신학회지(정보와통신), 31(11), 52-57.

네이버 쇼핑플랫폼. (2019.5.2). Retrieved from https://d2.naver.com/helloworld/1264836

정승윤 (2017). Factorization machine을 이용한 추천시스템 설계. 석사학위논문, 고려대학교 정보보호대학원.

조연선 (2019). 협업 필터링을 이용한 도서 추천 시스템이 학교도서관 이용에 미치는 효과 연구. 석사학위논문, 연세대학교 교육대학원.

홍유진 (2019). 도서 추천을 위한 학부생 대출데이터 기반의 연관 규칙 도출. 석사학위논문 고려대학교 교육대학원

Chollet, F., (2017), Xception: Deep Learning with Depthwise Separable Convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Honolulu, HI.

He, K., Zhang, X., Ren, S., & Sun, J., (2016), Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV.

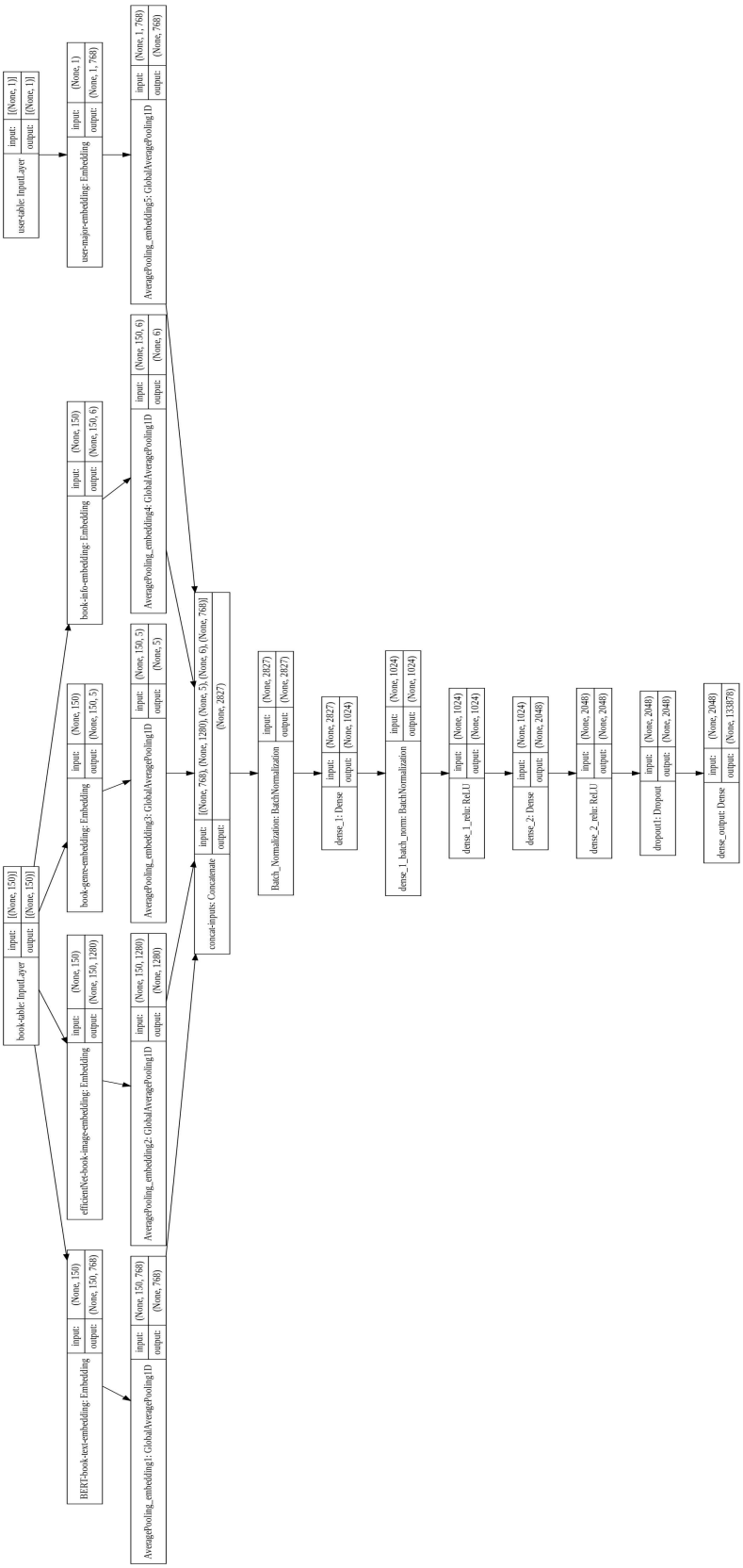
Paul,C., Jay,A.,& Emre,S. (2016, September), Deep Neural Networks for YouTube Recommendations. Paperpresented at the meeting of the RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems, Boston , MA.

Tan, M., & Le, Q. (2019, June). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA.

Yinhan, Liu., Myle, O., Naman, G., Jingfei, D., Mandar, J., Danqi, C., Omer, L., Mike, L., Luke, Z., & Veselin, S. (2019, July). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Retrieved from https://arxiv.org/abs/1907.11692v1

Yong, T., Hazarike, D., Poria, S., Cambria, E. (2018). Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 13(3), 55-75

[별첨 1] 다중 입력 모델의 기본 아키텍처



[별첨 2] 도서 제목 임베딩 코사인 유사도

[별첨 3] 최종 모델링에 사용된 3가지 변형 모델

	Model1	Model2	Model3
Batch Size	64		
Learning Rate	0.01		
Epoch	250		
Variable Training	title, image = False, genre = True		
Layer	Embedding	Embedding	Embedding
	Average Pooling	Average Pooling	Average Pooling
	Batch Normalization		
	Concatenate	Concatenate	Concatenate
	Batch Normalization	Batch Normalization	Batch Normalization
	Flatten	Dense	Dense
	Dense	Batch Normalization	Batch Normalization
	ReLU	ReLU	ReLU
	Dense	Dense	Dense
	ReLU	ReLU	ReLU
	Dropout	Dropout	Dropout
	Dense	Dense	Dense
Optimizer	Adagrad	SGD	SGD(momentum = 0.8)

[별첨 4] 모델의 도서 예측 목록과 사용자 검증결과 예시

순위	SVD	SGD	Collaborative Filtering	Deep Learning
1	여행의 이유	(사물인터넷의)빅데이터 개론	난중일기	The houses of Louis Kahn
2	82년생 김지영	(신)종합미생물학	(에밀 뒤르켐의) 자살론	Urban street design guide
3	(2018) 김유정 문학상	대통령의 글쓰기	인생에 한 번은 나만을 위해	Structural analysis and design of tall buildings
4	여자 둘이 살고 있습니다	우주의 7가지 놀라운 신비	이것이 불교의 핵심이다	Conceptual chemistry
5	피프티 피플	한국언론의 신뢰도	Kreyszig 공업수학	Eco skyscrapers
총점	4	4	4	3