PART. I
연구과제

네트워크 임베딩을 활용한 도서 및 친구 추천 시스템 설계

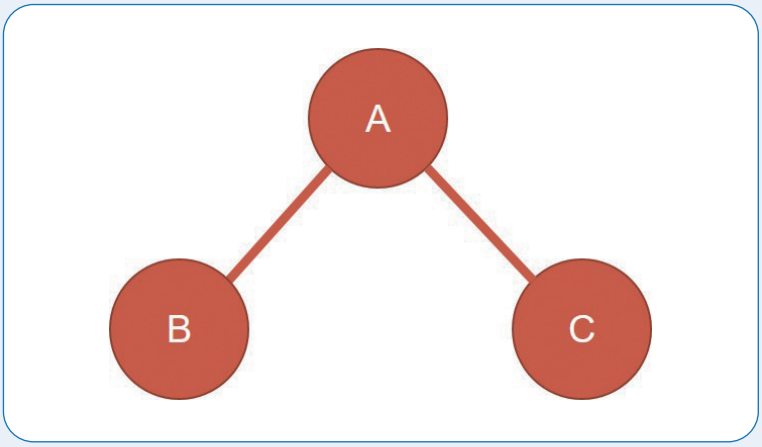
연구배경
연구내용
연구결과
향후계획

네트워크 임베딩을 활용한 도서 및 친구 추천 시스템 설계

숭실대학교 스마트시스템소프트웨어학과
윤진혁

연구과제

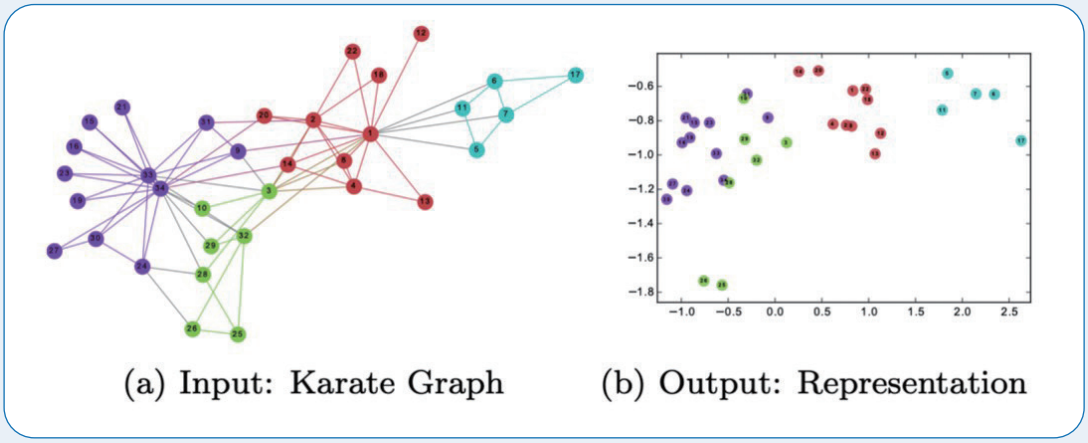
네트워크 구조는 추상적이고 복잡한 관계와 상호작용을 데이터로 표현하는 데에 적절하다. 또한, 복잡한 문제를 더 간단한 표현으로 단순화하고 반대로 관계의 종류를 정의하여 다양한 요소들의 상호작용을 컴퓨터가 처리 가능한 형태로 만들 수 있다. 하지만 네트워크의 경우 유클리드 공간으로 표현할 수 없으며 무한대의 차원을 가지는 프랙털 구조를 가진다.



[그림 1] 네트워크 구조

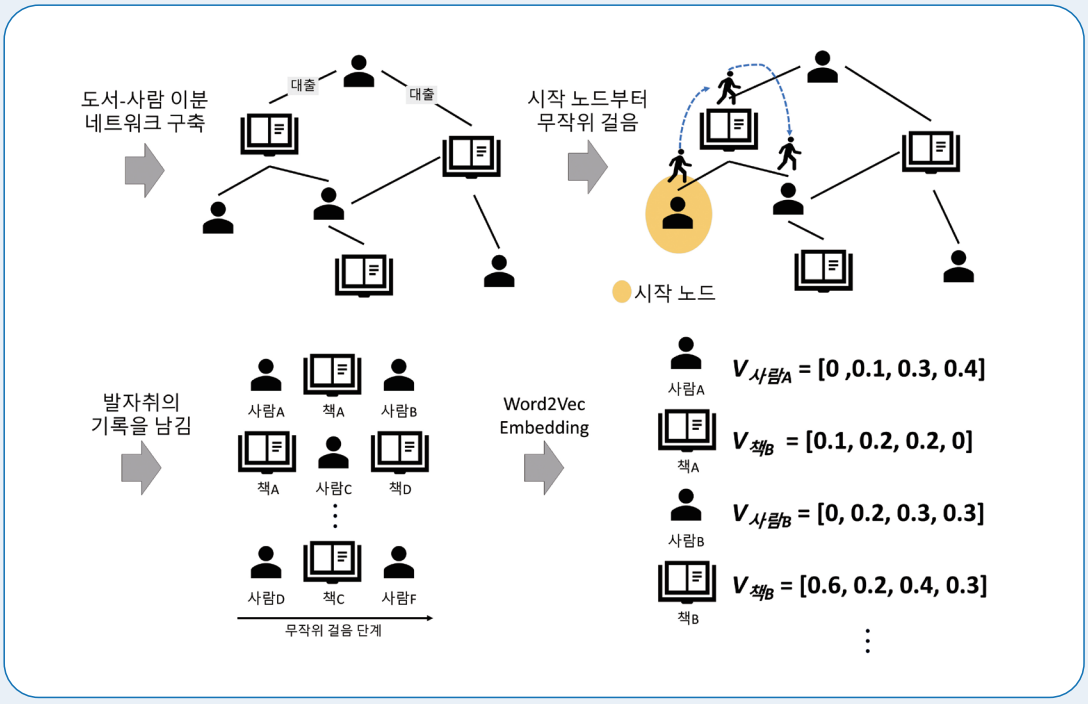
또한 네트워크의 경우 직접적인 연결이 없는 경우에 대한 관계 추정이 어렵고 무관계함을 가정하나, 실제로는 샘플링 문제 등으로 관측되지 않은 링크나 아직 만들어지지 않았으나 차후에 연결이 생길 가능성이 있는 노드쌍이 존재할 수 있다. 네트워크의 희소성으로 인하여 소수의 노드만 연결이 많이 존재하고, 대부분의 노드는 실제 정보가 적어 올바른 정보를 유추하는 것이 어려울 때가 많다.

이러한 단점을 보완하고자 Graph Neural Network(GNN)과 Graph Embedding의 개념이 도입되었다. 한 예로 Recurrent GNN은 한 노드의 feature와 그 이웃한 노드와 상호 간 링크의 현재 feature를 정의하고 이를 재귀적으로 업데이트하여 규칙성이 없는 Graph를 벡터 공간상에 표현하는 형태로 학습한다.



[그림 2] Graph Embedding

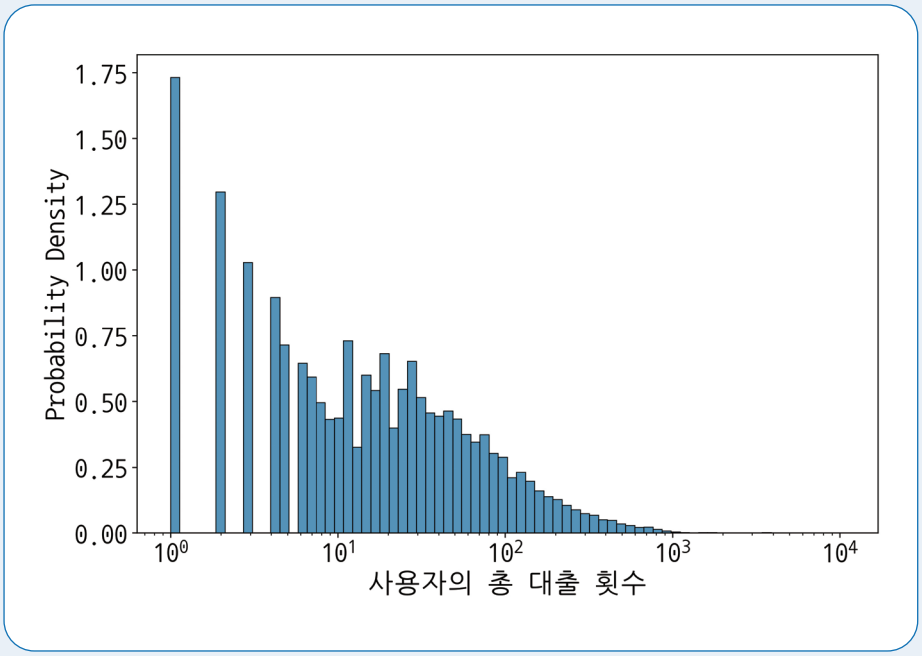
본 연구에서는 도서관 대출 데이터를 활용하여 이분(도서-사람) 혹은 삼분(도서-사람-강의) 네트워크를 구축하고, 이를 Random Walker를 사용한 방식으로 Embedding하여(그림 3) 도서관 서비스에 활용할 수 있는 방법론을 탐색하였다.



[그림 3] 도서-사람 이분 네트워크의 구조와 임베딩 방법

연구내용

먼저 실제 대출 기록에서 희소성 문제가 발생하는지, 두꺼운 꼬리 분포가 나타나는지를 확인하기 위하여 사용자별 대출 기록의 수 분포를 확인해 보았다. 대출 기록이 있는 사람 중 최소의 대출 횟수는 1회였으며, 최대 대출 사용자의 대출 건수는 10,613건으로 나타났으며 1인당 평균 4.38권을 대출하였고 표준편차는 126.11권으로 나타났다. 또한 대출 기록이 있는 책은 총 420,915권 (mms_id기준) 이었다. 대출 건수의 분포는 [그림 4]와 같이 나타났으며 꼬리가 두꺼운 형태의 분포가 나타났으나 그 형태가 멱함수 꼴을 나타내지는 않는 것으로 추정된다.



[그림 4] 도서관 이용자 1인의 도서 대출 수의 분포

다음으로 [그림 3]의 과정으로 Embedding한 사용자 및 도서의 vector가 실제를 잘 반영하는지에 대한 분석을 진행하였다. word2vec model의 경우 embedding에 사용하는 parameter에 따라 결과가 다르게 나타날 수 있다. 본 연구에서는 sequence의 길이를 30으로, 한 노드당 20번의 random walker를 출발시켜 50차원의 벡터 공간에 embedding하였다. embedding을 위해서 길이 5의 word window를 사용하였고, skip-gram with negative sampling 모델로 계산하였으며 5번의 epoch 수행하여 최종 모델을 도출하였다. 재현성을 위해 embedding을 위한 무작위 함수의 seed는 4,090을 사용하였다.

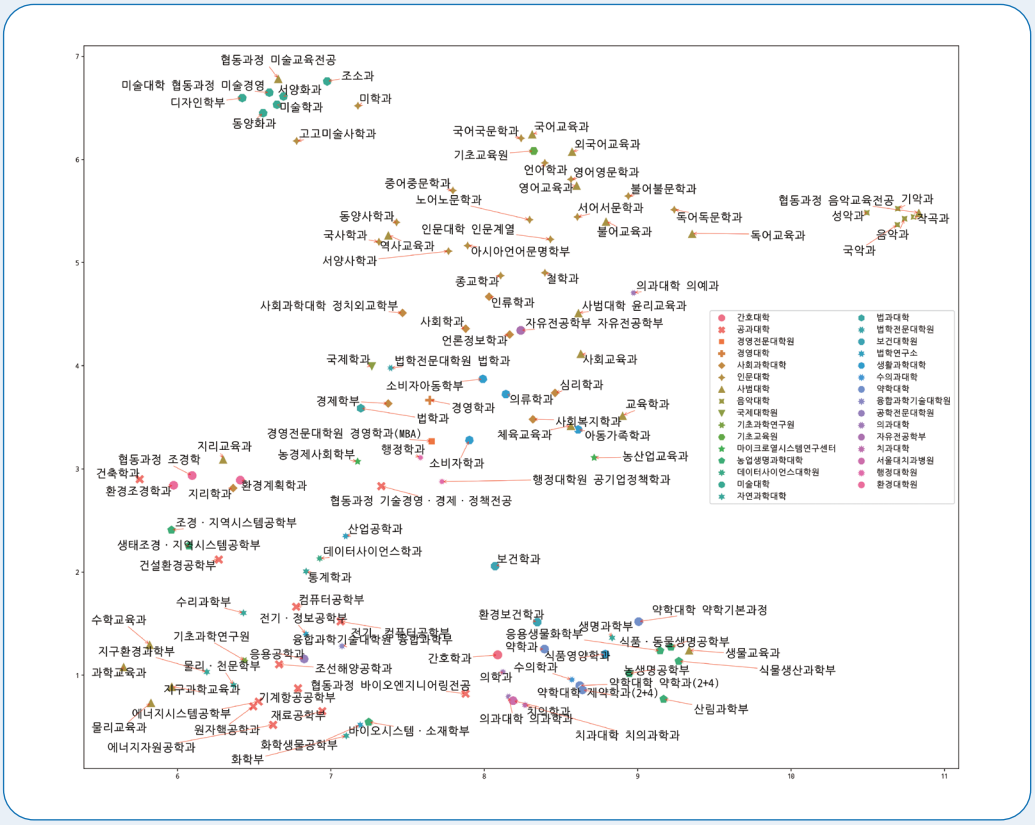
[그림 5]의 경우 위에서 구축한 이분 네트워크에서 구한 임베딩 벡터의 학과별 유사도를 2차원 공간에 표시하였다. 학과 벡터의 경우 학과에 속한 모든 학생의 벡터를 길이 1로 L2 정규화를 수행한 이후, 이렇게 정규화한 벡터를 평균하여 구하였다. 2차원 공간에 표시하기 위해 이렇게 구한 평균 벡터를 Uniform Manifold Approximation and

Projection (UMAP) 이라는 manifold learning 기법을 사용하여 차원 축소하였다.

[그림 5]에서 먼저 유사 학과가 근처에 위치하는 경향성을 확인할 수 있었다. 예를 들어 우측 상단에는 음악계통학과(협동과정 음악교육전공, 기악과, 성악과, 작곡과, 국악과) 등이 위치하였으며 좌측 상단에는 미술계통학과(협동과정 미술교육전공, 조소과, 미학과 등)가 위치함을 확인할 수 있었다.

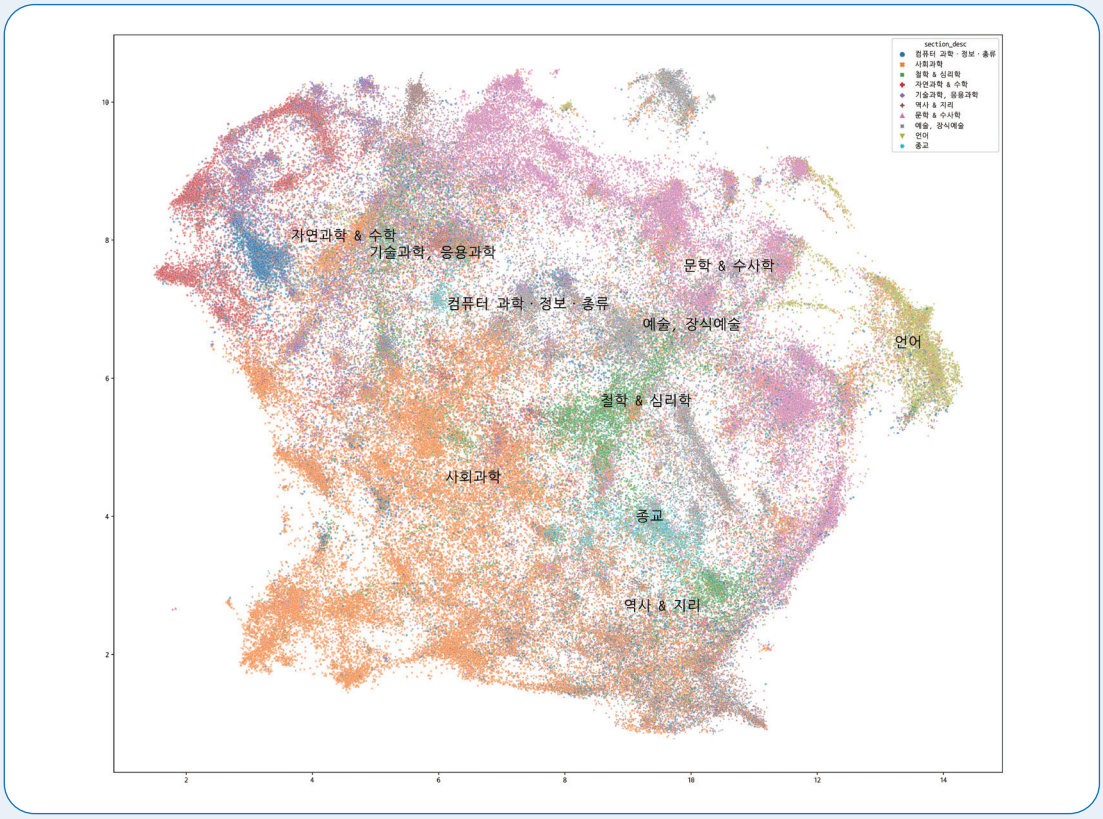
학과의 소속 단과대학보다는 유사 전공인 학과가 근처에 위치함을 확인할 수 있었다. 예를 들어 환경 및 건축 계통 학과(지리학과, 지리교육과, 환경계획학과, 환경조경학과, 건축학과 등)는 단과대학에 무관하게 모두 좌측의 유사 위치에 모여 있었으며, 유사하게 생명 및 의약학 관련 학과들(수의학과, 간호학과, 학과, 생명과학, 생물교육, 환경보건 등)은 모두 우측 하단에 위치한 것을 확인할 수 있었다.

교육 계통 학과들은 서로 모여있지 않고 전공이 유사한 학과와 비슷한 위치에 놓여있음을 확인할 수 있었다. 예를 들어 지구과학교육과는 지구환경과학부와, 물리교육과는 물리천문학부와 유사한 위치에 있었는데, 이를 통해 도서대출 패턴은 교육계통 학과와 전공이 유사한 학과가 비슷함을 확인할 수 있었다.



[그림 5] 이분 네트워크의 학과별 유사도. 학과의 유사도는 학과 학생의 normalized vector의 평균값을 사용하였으며, 시각화에는 데이터에 100명 이상 존재하는 학과만 사용하였다.

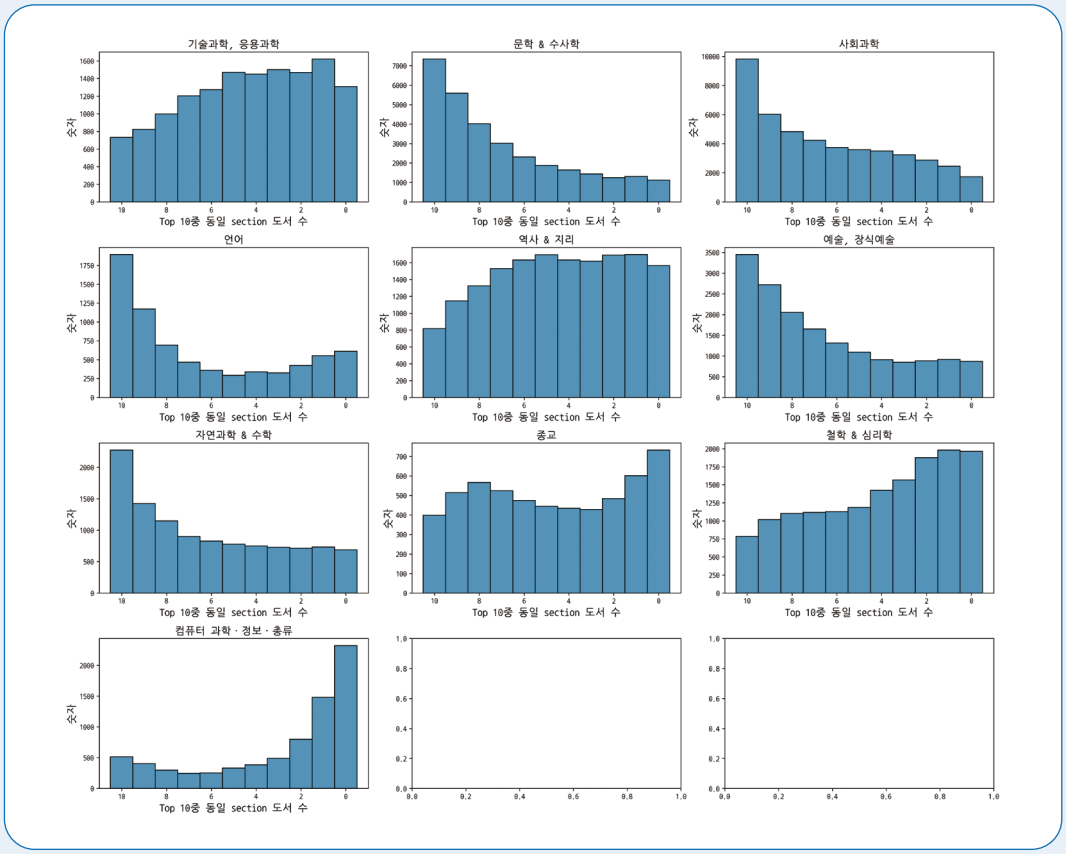
[그림 6]의 경우 위에서 구축한 이분 네트워크에서 구한 도서간 임베딩 벡터 유사도를 시각화 하였다. 역시나 2차원 공간에 표현하기 위하여 UMAP을 사용하였다. 각 도서의 색상은 도서관의 카테고리 분류 코드중 section code 10가지를 사용하였다. 또한 텍스트로 표기된 분야는 각 분야 도서 전체의 평균 벡터 위치에 두었는데, 개략적인 각 분야의 중심점 위치를 뜻한다. 아래 그림에서 유사한 섹션의 도서가 모여있는 것을 확인할 수 있는데, 이는 개인의 도서 대출 패턴은 비교적 일관적이며 본인이 관심있는 분야의 도서를 지속적으로 대출한다는 점을 뜻한다.



[그림 6] 이분 네트워크에서 구성한 도서별 대출 패턴 유사도

도서와 사용자 모두 고유한 vector를 가지기 때문에 이러한 vector를 기반으로 한 추천 시스템을 구축할 수 있다. 가장 단순하게는 자신과 가장 비슷한 vector를 가지는 아이템을 추천하는 방식이 있다.

[그림 7]에서는 cosine similarity를 기반으로 각 도서별 10개씩의 연관 도서를 추천받아 이들이 어떤 section에 분류되어 있는지 통계를 보여준다. 재미있게도 2가지 다른 패턴으로 구분되는 것을 관찰할 수 있는데, 문학&수사학, 사회과학, 언어, 예술&장식예술, 자연과학&수학의 경우 동일한 주제의 책을 더 많이 추천해주는 경향을 확인할 수 있었으나, 반대로 기술과학&응용과학, 역사&지리, 종교, 철학&심리학, 컴퓨터과학 관련 서적은 다른 분야의 책을 더 많이 추천해주는 것을 확인할 수 있었다.



[그림 7] 이분 네트워크에서 도서 1개를 기반으로 타 도서를 추천하였을 때 동일 section의 도서가 추천되는 경우의 수

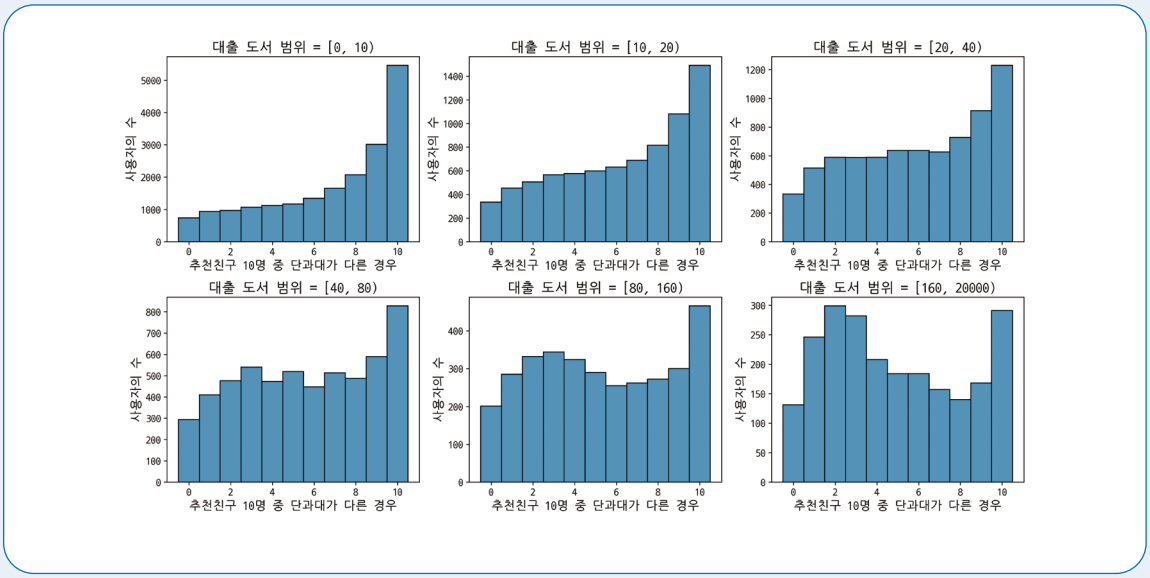
다음으로 추천시스템이 어떠한 사용자들에게 가장 가치가 높은지에 대해서 판단하기 위하여 추천시스템에서 추천해 준 도서 중 실제로 대출한 기록이 없는 도서가 몇 권이나 있나를 확인해 보았다.

[그림 8]에서 확인할 수 있듯, 대출한 도서의 수가 늘어날수록 추천 도서 중 더 많은 수의 도서가 기존에 대출 기록이 있는 도서로 나타났다. 하지만 80권 이상 160권 미만을 대출한 도서관을 자주 사용하는 사람의 경우에도 추천 도서 중 대출 기록이 없는 도서가 한 개라도 나타나는 경우가 대출 기록이 있는 도서들만 추천하는 경우보다 많았다. 160권 이상을 대출한 최상위 사용자의 경우 거의 모든 도서가 기존 대출 기록이 있는 도서였으며, 극소수의 미대출 도서만 추천된다는 것을 확인할 수 있었다.

추천 도서의 수를 늘린다면 이러한 고빈도 사용자에게도 도서를 추천할 수 있으나 전체적인 유사도가 낮아져서 추천의 질이 떨어질 것을 추측할 수 있다.

위와 유사한 방식으로 cosine similarity 기반으로 본인의 독서 대출 취향과 가장 유사한 친구를 추천해주는 시스템도 설계가 가능하다. [그림 8]은 가장 유사도가 높은 10명의 사용자를 추천해주는 경우 동일 단과대를 추천받는 경

우와 다른 단과대를 추천받는 경우를 사용자의 도서 대출 수에 따라 그린 그림이다. 도서의 대출 수가 늘어날수록 단과대가 같은 학생을 추천하는 비율이 높아지는 것을 알 수 있다. 도서 대출 수가 많은 경우 같은 단과대를 추천하는 비율이 올라가는 것은 도서관을 전공 관련 서적을 대출하는데에 주로 사용하는 학생의 비중이 높기 때문으로 추측된다. 동시에 추천받은 10명 모두가 단과대가 다른 경우의 비율은 일정 수준 이상으로 유지되는 경향도 보이는데, 이는 일정 비율의 사용자는 전공서적보다는 교양서적이나 기타 서적을 대출하기 때문으로 추정된다. 다만 위의 추정을 뒷받침하기 위해서는 도서가 교양서인지 전공서인지를 구분해야 하는데, 이는 현재 데이터에는 구분되어 있지 않아 파악이 어렵다.



[그림 8] 도서 대출 수에 따라 이분 네트워크 모델에서 추천한 친구 10인 중 단과대가 같은 경우와 다른 경우. 도서 대출의 수가 늘어날수록 같은 단과대를 추천하는 비율이 늘어나는 것을 볼 수 있다.

연구결과

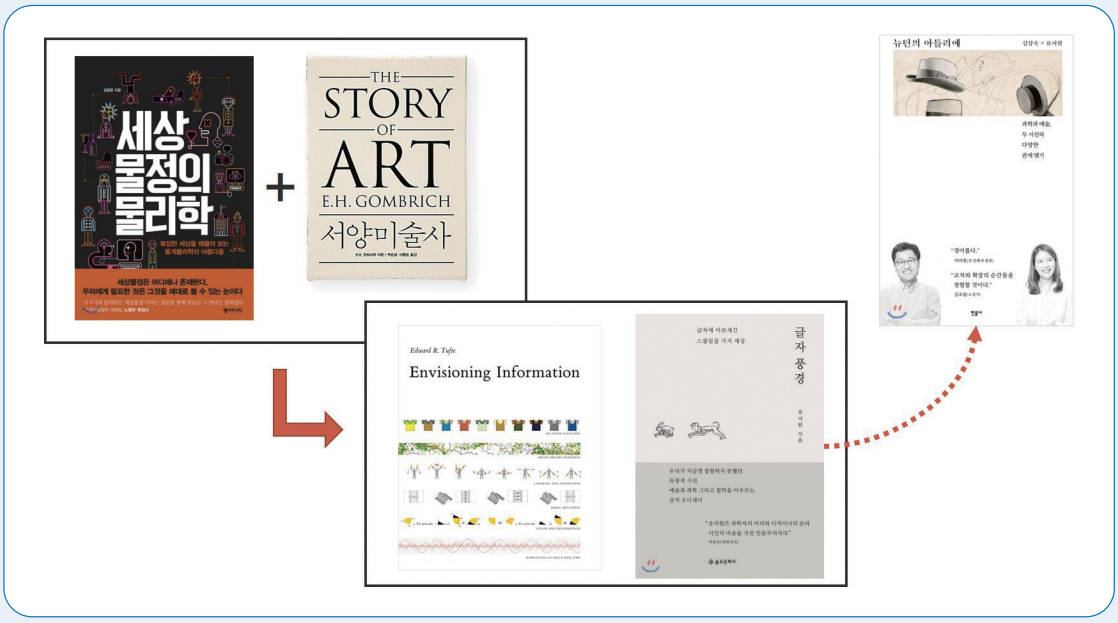
제안분석 모델 1 - 도서 대출 기록을 활용한 추천 도서 시스템

위에서 embedding한 데이터를 응용하면 도서 대출 기록이나 본인이 입력한 책을 기반으로 도서를 추천해 주는 시스템을 구축할 수 있다. 먼저 입력한 책들을 바탕으로 도서를 추천하는 시스템을 제작해 보았다. 입력한 도서들의 벡터의 길이를 동일하게 만드는 L2 정규화를 수행한 후 평균 벡터를 구하였고, 이 평균 벡터와 가장 유사한 방향성을 가지는 다른 책을 추천할 수 있다. 이 시스템이 제대로 작동하는지에 대한 예시를 찾기 위하여 복잡계 물리학의 대표적인 도서 3권을 입력으로 받아 추천 도서를 찾는 예시를 보았다. (그림 9)

[그림 9] 동일 분야로 입력하는 경우 도서 추천 예시 : 복잡계 네트워크 관련 대표 도서 3가지(세상물정의 물리학, 구글신은 모든 것을 알고 있다. 버스트 : 인간의 행동 속에 숨겨진 법칙)를 입력하는 경우의 추천 도서. 대부분이 복잡계 네트워크를 다루고 있다.

[그림 9]의 예에서 3가지의 유사 도서를 입력하는 경우, A.L. Barabasi, M. Buchanan, S. Strogatz 등이 쓴 복잡계 물리학에 대한 다양한 저서가 주로 추천되는 것을 확인할 수 있었으며 이 방법론으로 도서의 주제를 잘 찾을 수 있음을 확인하였다.

이를 확장하여 전혀 다른 주제의 도서 두 가지를 입력하는 경우도 확인해 보았다. 복잡계 물리학 저서인 〈세상 물정의 물리학〉과 서양미술사의 가장 대표적 교과서인 〈곰브리치 서양미술사〉 두 가지 책을 입력으로 사용한 결과에서는(그림 10) 정보를 효율적이고 심미성있게 시각화하는 방법을 다루는 E. R. Tufte의 〈Envisioning Information〉과 타이포그래피에 대해 다루는 유지원 교수의 〈글자풍경〉 두 권이 가장 높은 유사도를 보였다.



[그림 10] 도서 추천 예시 2: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사)를 동시에 입력하는 경우의 추천 도서 예시

2위로 추천된 〈글자풍경〉의 작가인 유지원 교수의 경우 미술 전공이나 〈글자풍경〉의 출간 이후 경희대 물리학과의 김상욱 교수와 함께 〈뉴턴의 아틀리에〉 라는 교양 과학 서적을 출간하였다. 이러한 예시를 통해 임베딩을 통한 도서 추천이 어느 정도 사용자 취향에 기반한 도서 추천이 가능하게 함을 확인할 수 있다.

또한 이 모델은 단순히 도서만 입력할 수 있는 것이 아니라 사람의 정보도 같이 입력할 수 있는데(그림 11) 동양사학과 대학원생을 위의 〈곰브리치 서양미술사〉 및 〈세상물정의 물리학과〉 동시에 입력하는 예시를 보면 단순히 책을 입력하는 것과 다르게 사용자의 전공에 기반하여 역사에 관련된 도서를 추천하는 비율이 크게 늘었음을 확인할 수 있다. 이러한 도서-사용자 동시 임베딩을 통하여 도서를 추천하는 경우 텍스트 등을 통해 단순히 도서의 내용 기반으로 추천하는 것 보다 사용자의 취향에 더 잘 맞는 추천을 해 줄 가능성이 높으며, 도서관 만족도를 향상시킬 수 있다.

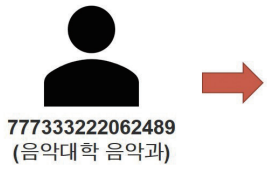
```
# 테스트
# 99391301702591: 세상물정의 물리학
# 99134781502591: 곰브리치 서양미술사
# 777333222026412: 동양사학과 대학원생
input_list = ["99391301702591", "99134781502591", "777333222026412"]
get_most_similar_books_from_multibook(input_list)[["title", "author"]]
```

	title	author
12	조선의 무기 /	강신엽.
8	鹽鐵論 : 소박하면 풍족해지고, 사치하면 기근이 온다 /	Huan, Kuan,
6	(시장으로 나간) 조선백자 : 분원과 사기장의 마지막 이야기 /	박은숙,
11	발해 대외관계사 자료 연구 /	장재진
10	고대 동북아시아 교통사 /	Wang, Mianhou
1	(옛날 이야기처럼 재미있는) 곰브리치 세계사 /	Gombrich, E. H.
0	군주론 /	Machiavelli, Niccol?,
13	고려의 북진정책사 /	장학근.
9	(신편) 발해국지장편 /	Jin, Yufu,
4	고려, 북진을 꿈꾸다 : 고구려 영토 회복의 꿈과 500년 고려전쟁사 /	정해은,

[그림 11] 도서 추천 예시3: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사), 동양사학과 대학원생을 입력하는 경우의 추천 도서 예시. 위의 단순 추천과 책 목록이 달라져 역사학 관련 도서가 증가하였음을 볼 수 있다.

제안분석 모델 2 - 도서 대출 기록을 활용한 독서 모임 친구 추천 시스템

임베딩 모델에서는 도서 이외에도 사용자도 모두 벡터가 주어져 있으므로 사용자들의 독서 취향에 맞는 친구를 추천해주는 시스템도 구축할 수 있다. [그림 12]는 임의로 선정된 음악대학 대학원생과 가장 유사한 친구 10인을 추천하는 경우를 예시로 들었다. 이런 경우 유사한 전공의 사용자가 주로 추천이 되나, 공과대학이나 사범대학 소속의 대학원생 또한 추천이 됨을 확인할 수 있다. 학과 기준이 아닌 독서 취향을 통해 친구를 추천하므로 단순히 전공을 넘어 본인과 관심사를 공유하는 친구를 추천해줄 수 있다.



	user_id	group	college	department
14	777333222108498	대학원생	음악대학	음악과
3	777333222056278	대학원생	사범대학	협동과정 음악교육전공
15	777333222114036	학부생	음악대학	국악과
10	777333222075609	대학원생	음악대학	음악과
12	777333222077357	대학원생	공과대학	전기·정보공학부
17	777333222119068	대학원생	음악대학	음악과
5	777333222062407	대학원생	음악대학	음악과
19	777333222131976	학부생	음악대학	국악과
22	777333222190484	강사	음악대학	국악과
18	777333222122354	학부생	음악대학	국악과

[그림 12] 친구 추천 예시: 음악대학 대학원생을 입력하는 경우 유사 학생들이 주로 검색된다.

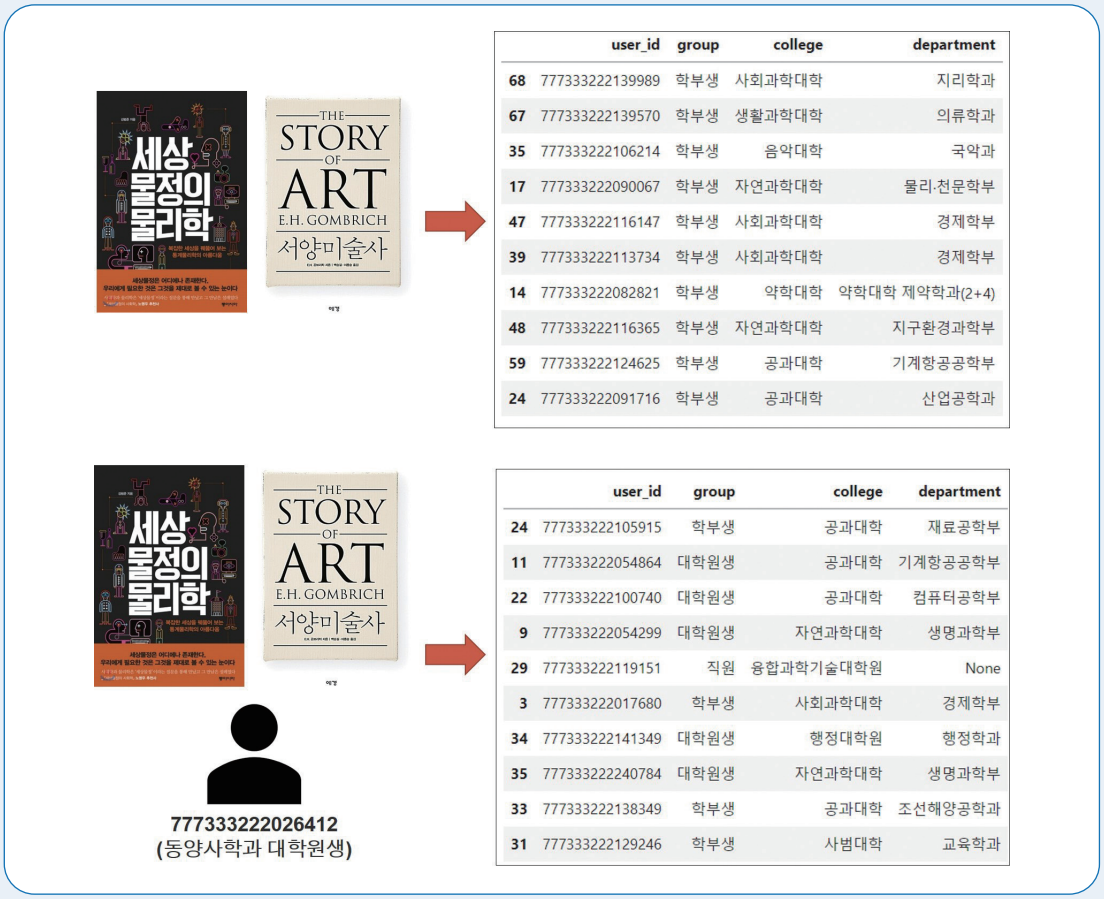
또한 이 모델은 도서의 정보를 입력하여 도서와 관련있는 사용자를 추천해줄 수도 있다. 예를 들어 도서관에서 분야별 추천 도서 목록을 선정하여 관심있는 사용자에게 발송하는 경우 등에 사용할 수 있다. 제안분석모델 1에서 사용된 복잡계 물리학 관련 3가지 도서를 입력하는 경우를 다시 고려해 보자(그림 13). 3개의 도서는 모두 물리학자가 집필한 물리학 관련 도서로 볼 수 있다. 하지만 이러한 도서에 관심을 보일 것으로 예상되는 사람들은 물리학과나 자연과학 대학 이외에도 사회학, 인문대학, 융합과학기술대학원 등 다양한 전공을 가지고 있음을 확인할 수 있다. 이는 사회학이나 경제학 등에서도 복잡계 네트워크 관련 기법을 사용하는 연구가 다수 존재하므로 이러한 분야적 관심도에 기반한 것이라 추정된다.

```
# 테스트
# 99391301702591: 세상물정의 물리학
# 99127554202591: 구글신은 모든 것을 알고 있다
# 99172392802591: 버스트 : 인간의 행동 속에 숨겨진 법칙
input_list = ["99391301702591", "99127554202591", "99172392802591"]
get_most_similar_users_from_multiitem(input_list)[["user_id", "group", "college", "department"]]
```

	user_id	group	college	department
19	777333222138625	학부생	자유전공학부	자유전공학부 자유전공학부
12	777333222090067	학부생	자연과학대학	물리·천문학부
16	777333222116147	학부생	사회과학대학	경제학부
18	777333222138028	학부생	자연과학대학	물리·천문학부
26	777333222190626	대학원생	사범대학	사회교육과
11	777333222086661	대학원생	공과대학	재료공학부
14	777333222109767	대학원생	융합과학기술대학원	None
15	777333222113734	학부생	사회과학대학	경제학부
22	777333222173944	회원제	인문대학	협동과정 인지과학전공
13	777333222099700	학부생	공과대학	재료공학부

[그림 13] 친구 추천 예시 2: 복잡계 네트워크 관련 대표 도서 3가지(세상물정의 물리학, 구글신은 모든 것을 알고 있다, 버스트 : 인간의 행동 속에 숨겨진 법칙)를 입력하는 경우 이공계생 뿐 아니라 다양한 분야의 친구를 추천해준다.

위의 두 예시에서 사용된 단순 쿼리 이외에도 사용자-도서 복합 쿼리를 사용하여 관심도 있는 사람을 추천해줄 수 있다(그림 14). 예를 들어 독서 모임의 리더가 있고, 리더가 추천하는 독서 목록을 동시에 가지고 있는 경우 이 두 가지를 모두 고려하여 리더의 독서 취향 및 리더가 새로 읽고자 하는 책에 모두 관심 있는 사용자를 추천하는 것을 수행할 수 있다.



	user_id	group	college	department
68	777333222139989	학부생	사회과학대학	지리학과
67	777333222139570	학부생	생활과학대학	의류학과
35	777333222106214	학부생	음악대학	국악과
17	777333222090067	학부생	자연과학대학	물리·천문학부
47	777333222116147	학부생	사회과학대학	경제학부
39	777333222113734	학부생	사회과학대학	경제학부
14	777333222082821	학부생	약학대학	약학대학 제약학과(2+4)
48	777333222116365	학부생	자연과학대학	지구환경과학부
59	777333222124625	학부생	공과대학	기계항공공학부
24	777333222091716	학부생	공과대학	산업공학과

	user_id	group	college	department
24	777333222105915	학부생	공과대학	재료공학부
11	777333222054864	대학원생	공과대학	기계항공공학부
22	777333222100740	대학원생	공과대학	컴퓨터공학부
9	777333222054299	대학원생	자연과학대학	생명과학부
29	777333222119151	직원	융합과학기술대학원	None
3	777333222017680	학부생	사회과학대학	경제학부
34	777333222141349	대학원생	행정대학원	행정학과
35	777333222240784	대학원생	자연과학대학	생명과학부
33	777333222138349	학부생	공과대학	조선해양공학과
31	777333222129246	학부생	사범대학	교육학과

[그림 14] 친구 추천 예시 3: 복잡계 및 통계물리 서적(세상물정의 물리학)과 미술 서적(곰브리치 서양미술사)의 도서 조합을 입력하는 경우와 동양사학과 대학원생을 함께 입력하는 경우의 추천 도서 예시 비교. 아래 결과의 경우 위의 단순 추천과 다른 사람을 추천하는 것을 알 수 있다.

제안이유

코로나19 이후 비대면이 일상이 되며 도서관 서비스에 대한 주된 수요가 단순한 도서 대출에서 다양한 정보 서비스로 개편되고 있다. 기존에는 동아리나 학생회 등 다양한 경로로 본인과 유사한 취향을 가진 사람을 찾는 방법이 있었으나, 코로나19 이후 비대면 시대에 입학한 2020학번 이후의 학생들은 사회적 거리두기 등으로 위의 활동이 거의 불가능하였다. 도서관에서 취향에 맞는 도서를 추천하고, 더 나아가 동일 취향을 공유하는 독서 모임 등을 구성한다면 비대면 시대의 새로운 친구 찾기 경로가 될 수 있다. 독서 모임의 경우 서로 독립적으로 도서를 읽은 이후 서로 의

견을 교환하는 짧은 모임이 이어지는 형태로, 취향이 잘 맞는 친구를 구한다면 대면/비대면에 무관하게 높은 친밀도를 형성할 수 있다는 장점이 있다. 위에서 제시한 추천 시스템 등을 활용하면 대면/비대면 독서 모임 등을 구성하여 도서관이 단순히 도서를 대여하는 수동적 역할 뿐 아니라 비대면 시기에 교내의 결속력을 다지는 적극적인 매개체 역할을 할 수 있을 것으로 보인다.

향후계획

- **도서관 홈페이지의 추천 도서 서비스 개편에 활용**
- **메일링 리스트 등을 통한 사용자 추천 도서 발송**
- **1학년 논술 등 기초 과목 등과 연계하여 학생의 취향에 맞는 도서 추천으로 도서관 이용률 및 만족도 향상**
- **단순 대출량 기반이 아닌 고도화된 서비스를 통해 도서관의 도서 구매시 기반 데이터로 활용: 학과별 관심 도서 분야 등을 추출하여 도서 구매시 선 반영**