Movie Recommender System 프로젝트 보고서

2023. 09. 05.
Codestates AI 18기

무마카세 팀
류재영, 남자인, 권구현, 부지환

목차

1.프로젝트 개요

1.1 추진배경 및 기획 목적

1.2 프로젝트 목표

1.1 추진 배경 및 기획 목적

- 영화 추천 시스템
- 고객 만족도 증가를 통해 고객 이탈을 최소화
- 영화 추천 서비스의 고도화

1.2 프로젝트 목표

프로젝트 구현 방법

목표 1 : 추천 시스템 성능 개선

- **비개인화 추천 시스템**
  - Steam Rating 공식 적용
- **개인화 추천 시스템**
  - 모델 성능 고도화 : KNN
  - 모델 구조 변경 : Hybrid
  - 다른 모델 활용 : 회귀모델, GNN

    SVD, NMF, SASRec, BERT4Rec

목표 2 : 평가 방안 수립

- **Regression**
  - MAE, RMSE
- **Classification**
  - Precision@k
  - Recall@k
  - Hit@k
  - NDCG@k

1.2 프로젝트 목표

팀 구성 및 담당 역할

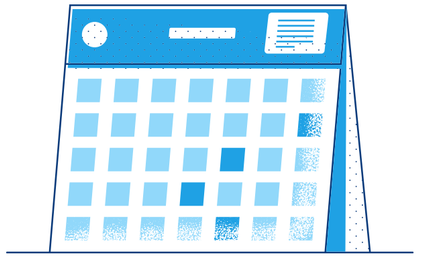
이름	담당 업무	공동 업무
류재영	• Sequential 모델 구축	• EDA • 비개인화 추천 시스템
남자인	• 하이브리드 모델 적용	
권구현	• 협업 필터링 및 GNN 모델 적용	
부지환	• 콘텐츠 기반 필터링 모델 적용	

- **운영체제** : Windows 11/ Mac
- **버전 관리** : Git/ Github
- **개발 언어** : Python 3.10.

– Sklearn, Tensorflow, Surprise, Recbole

1.2 프로젝트 목표

프로젝트 일정



진행 과정	8월 둘째 주	8월 셋째 주	8월 넷째 주	8월 다섯째 주	9월 첫째 주
도메인 조사 및 베이스 코드 이해					
추천 시스템 구축					
평가지표 적용					
결과 정리					
발표 준비					

2. 프로젝트 수행 방법

2.1 추천 시스템 개요

- 사용자의 정보 데이터를 분석, 개인의 취향에 맞는 아이템을 추천하는 알고리즘
- 고객에게 다양한 아이템을 추천함으로써 고객의 정보 데이터가 누적되고, 이를 통해서 고객의 취향과 니즈를 파악할 수 있음

표1 추천 목적의 네 가지 유형과 사례

목적	설명	예시	측정지표
Best Recommendation	잘 팔리는 상품 추천	베스트셀러	인기 순위 비교
Related Recommendation	관심 있는 상품과 유사한 상품 추천	이 상품을 조회한 고객이 구매한 상품	아이템 간 유사도
Personalized Recommendation	나에게 개인화된 상품 추천	'홍길동' 님만을 위한 추천	사용자의 아이템 선호도
Context Aware Recommendation	상황에 적합한 상품 추천	봄이 되면 '벚꽃엔딩', 비가 오면 '비처럼 음악처럼'	상황과 아이템의 유사도

출처 : : https://dbr.donga.com/article/view/1203/article_no/8734

2.2 데이터 셋

2.2.1 데이터 셋 내용 설명

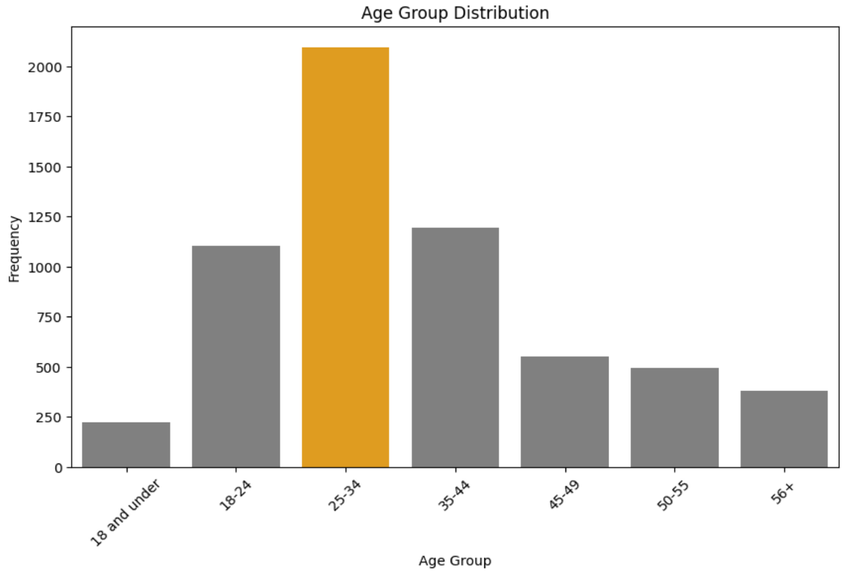
MovieLens 1M 데이터 셋

- 총 3 개의 데이터 파일 : **Movies, ratings, users**
- Movies : 영화 id, 제목 (연도), 장르
- Ratings : 사용자 id, 영화 id, 평점, 타임스탬프
- Users : 사용자 id, 성별, 나이, 직업코드, 우편번호

2.2 데이터 셋

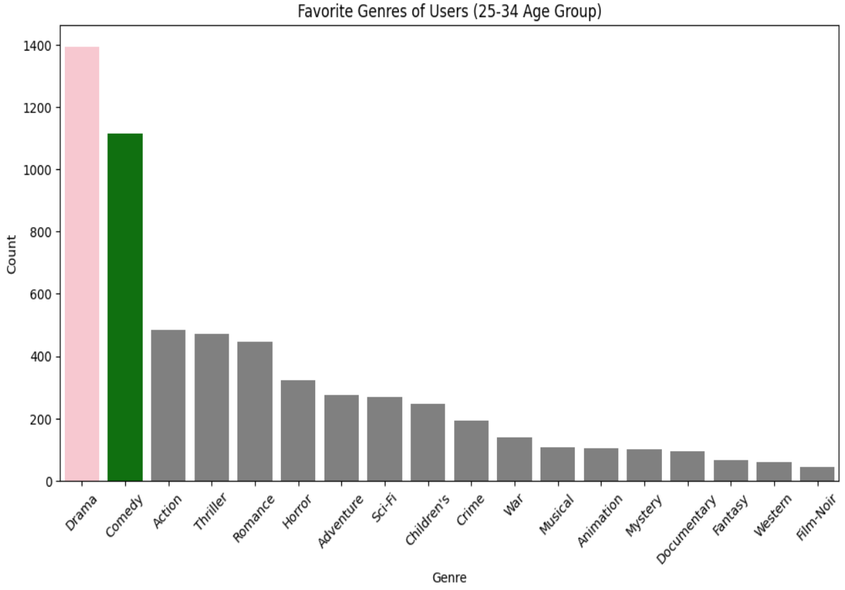
2.2.2 EDA 시각화

- 시청자의 나이 및 장르 분포



유저의 시청 연령 분포

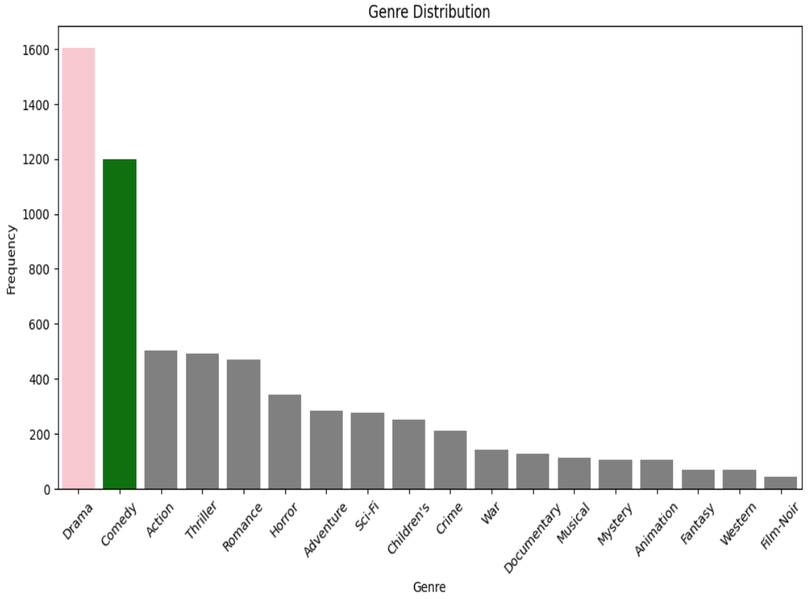
- 주 사용자 : 25-34세



주 사용 연령대의 장르 분포

1위 : 드라마

2위 : 코미디



전체 장르 분포

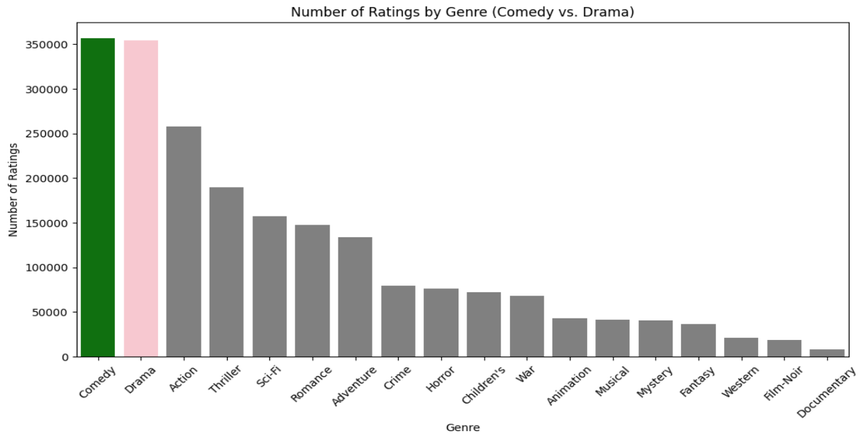
1위 : 드라마

2위 : 코미디

2.2 데이터 셋

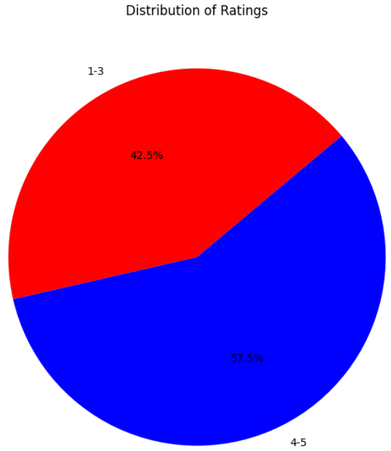
2.2.2 EDA 시각화

- **평점 분포**



장르별 별점 개수

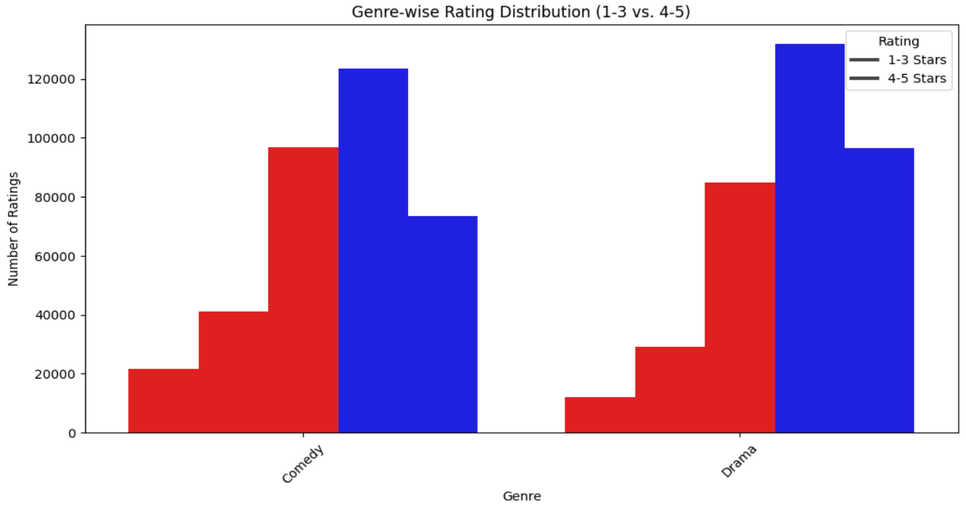
코미디 > 드라마



전체 평점 분포

1-3점 : 42.5 %

4-5점 : 57.5 %



주요 장르별 평점 분포도

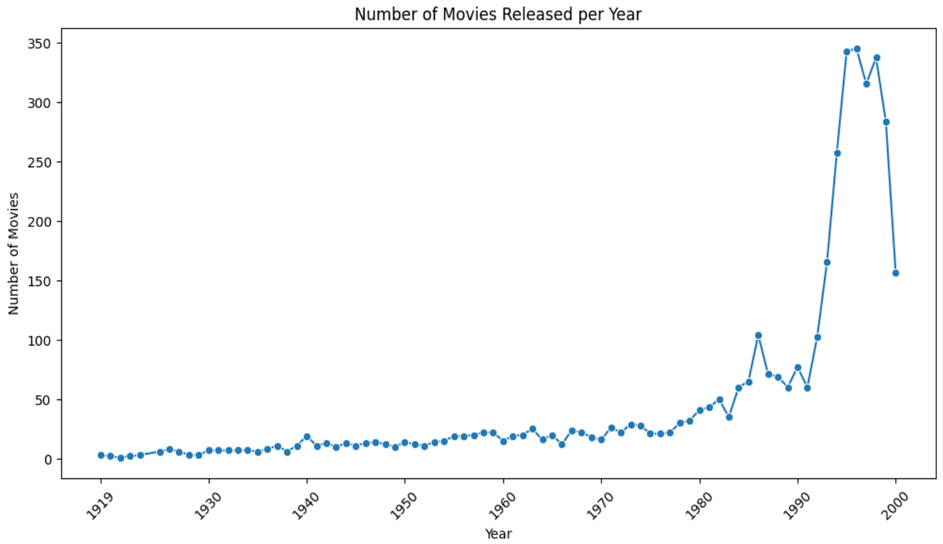
드라마가 코미디보다 긍정적인 반응을 보임

2.2 데이터 셋

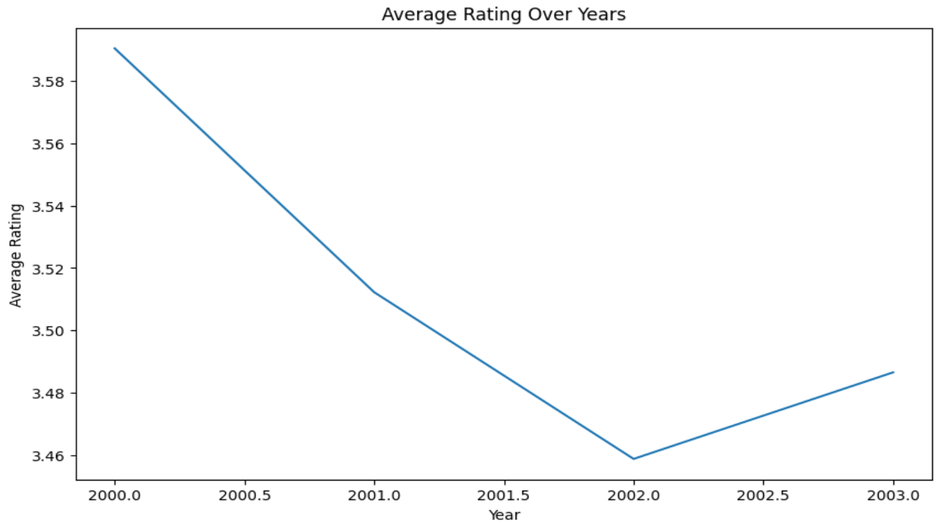
2.2.2 EDA 시각화

- 연도별 추세

연도별 개봉 영화 수



연도별 평균 별점 경향



1990~2000사이 영화 제작이 가장 활발하여 다양하고 재밌는 영화는 많이 나왔으나,

지금의 추천 서비스에 부족한 점이 많아 취향에 맞지 않은 영화를 본 결과 평균 별점은 많이 낮아졌다.

2.2 데이터 셋 설명

2.2.2 EDA 시각화

- **EDA Case Study : USER 1**

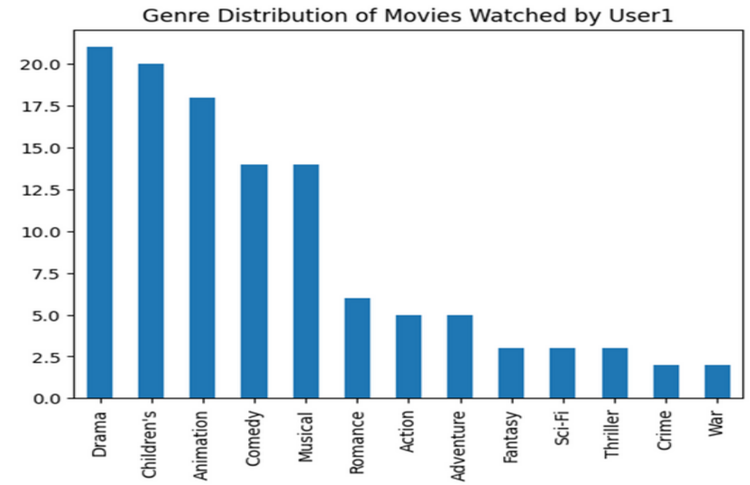
User 정보 : 여성 / 18세 미만 / 학생

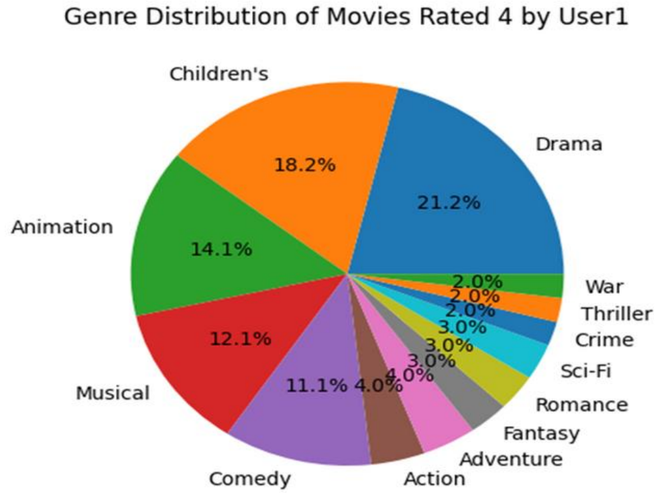
- **총 시청 영화 개수 : 53 개**
- Drama / Children' s / Animation / Comedy / Musical 위주로 시청함.
- 4점 이상으로 재밌게 본 영화비율도 상동
- Action, Romance, Crime, Thriller 등의 영화도 종종 4점 이상을 주는 경향이 있음.

▶ EDA 마무리

분석 결과를 토대로 User 1에 대한 각 모델의 추천 리스트를 뽑아 비교해볼 에정임

User1이 4점 이상의 평점을 남긴 장르의 분포 및 비율



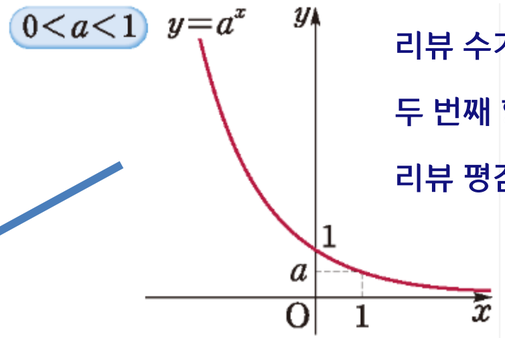


2.3 비개인화 추천 시스템

적용 알고리즘 개념

- **Steam Rating Algorithm**

$$Rating = ReviewScore - (ReviewScore - 0.5)2^{-\log(Totalreviews+1)}$$



리뷰 수가 많아질수록

두 번째 항은 0에 근접함에 따라

리뷰 평점을 그대로 반영 가능

(평균 긍정률) $ReviewScore = \frac{PositiveReviews}{TotalReviews}$ $TotalReviews = PositiveReviews + NegativeReviews$

$PositiveReviews$ = 4점이상의 영화
$NegativeReviews$ = 3점이하의 영화

리뷰 수를 반영하여 점수를 보정하는 알고리즘

리뷰 수에 비례하여 점수 반영 정도 증가 => 즉 같은 점수라도 **리뷰 수에 따라 점수를 보정**해 신뢰성을 확보함

기존 모델 : 인기도 기반 추천	적용 모델 : Steam Rating Algorithm
단순히 영화를 많이 시청한 순으로 나열	1 ~ 3 점은 0으로, 4 ~ 5점은 1로 점수를 조정해서 본 식에 적용 평균 평점이 0.5이하일 시 두 번째 항은 음수가 되어 점수를 높게 보정.

출처: https://steamdb.info/blog/steamdb-rating/

개인화추천시스템

-콘텐츠 기반 필터링-

발표자 : 부지환

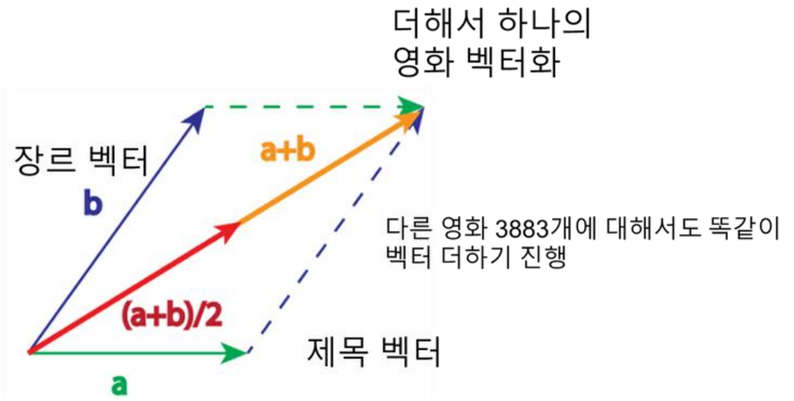
2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

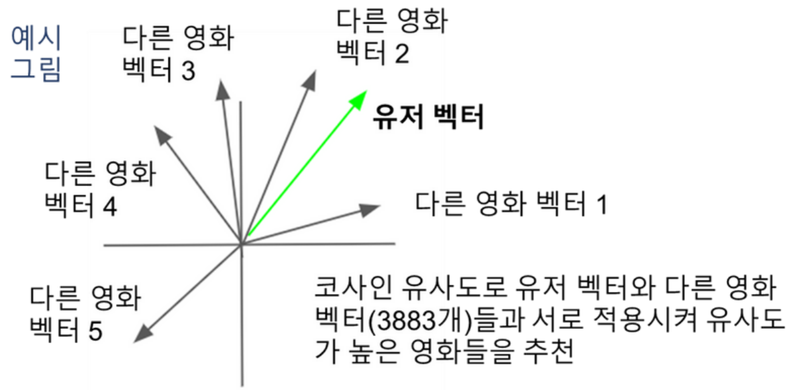
- **KNN 고도화 모델**

  - 기본 모델은 가까운 유사도를 적용하는데 “minkowski” 유사도와 알고리즘 자동 적용됨
  - 데이터 포인트 간의 대상이 되는 이웃 데이터에 대해 설정한 유사도에 따라 거리를 측정하는 알고리즘
  - 위 영화들에 대한 데이터는 각 제목 벡터와 장르 벡터를 더해서 하나의 벡터로 나타냄.
  - 모델은 코사인 유사도를 적용해 다른 유사도와 달리 벡터의 내적을 적용할 때 각 벡터의 길이를 또 나눠줌으로써 벡터의 크기와 상관없이 즉 방향성으로만 각 단어의 유사도를 봄.
  - 이 모델은 영화 벡터와 유저 벡터와의 유사도를 계산해 유저에 맞는 영화들을 추천해줄 수 있음.
  - 다음과 같이 영화 정보만을 기반으로 모델링 한것을 콘텐츠 기반 모델링이라 하고 유저 정보를 기반으로 모델링 한 것은 협업 기반 모델링이라 함.

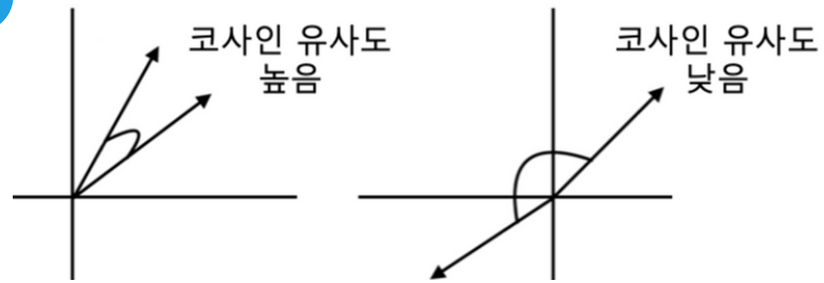
1



2



3



2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

- **회귀 모델**

- 회귀 모델을 기반으로 모델을 만듦.
- 이 때 해당하는 장르에 따라 장르를 컬럼화시킴.

	Western	Mystery	Sci-Fi	Comedy	Documentary	Crime	Action	Drama	War	Animation	Fantasy	Musical	Film-Noir	Adventure	Thriller	Children's	Romance	Horror
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0

- 여러 개 컬럼을 분할화 시킨 이유는 각 영화에 대해 학습을 시키고 각 유저가 다른 영화에 대한 학습을 진행하기 위함임.
- 유저 간의 유사도를 고려하지 않고 영화 데이터와 개별 유저를 집중하기 위함.

2.4 개인화 추천 시스템

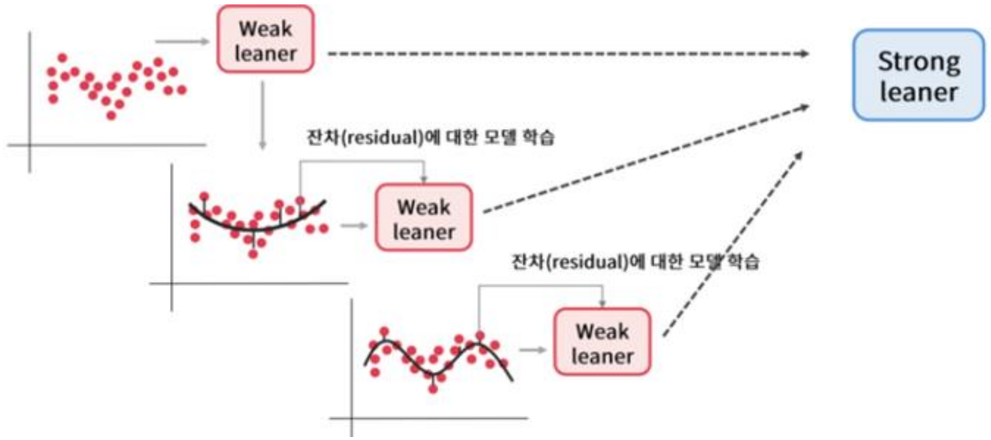
2.4.1 콘텐츠 기반 모델

1) Gradiant boosting

실제값과 예측값의 잔차의 기울기를 줄어드는 방향으로 학습시키고

이 때 학습은 여러 모델(학습기)등으로 순차적으로 적용시켜

학습해나가는 방식



2) Kneiborsregressor

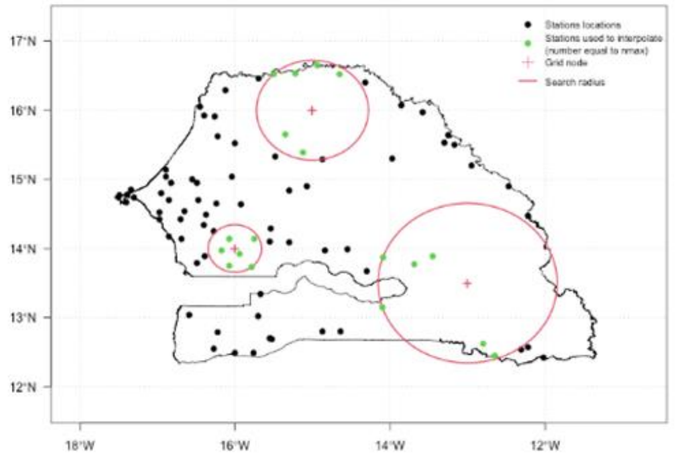
해당 모델은 knn 기반인 회귀 모델임.

knn으로 훈련된 데이터 셋에서 최대 이웃의 개수의 설정에 따라

지역 보간법을 실행

이 때 이웃의 개수에 따라 공간적인 반경을 생성하여

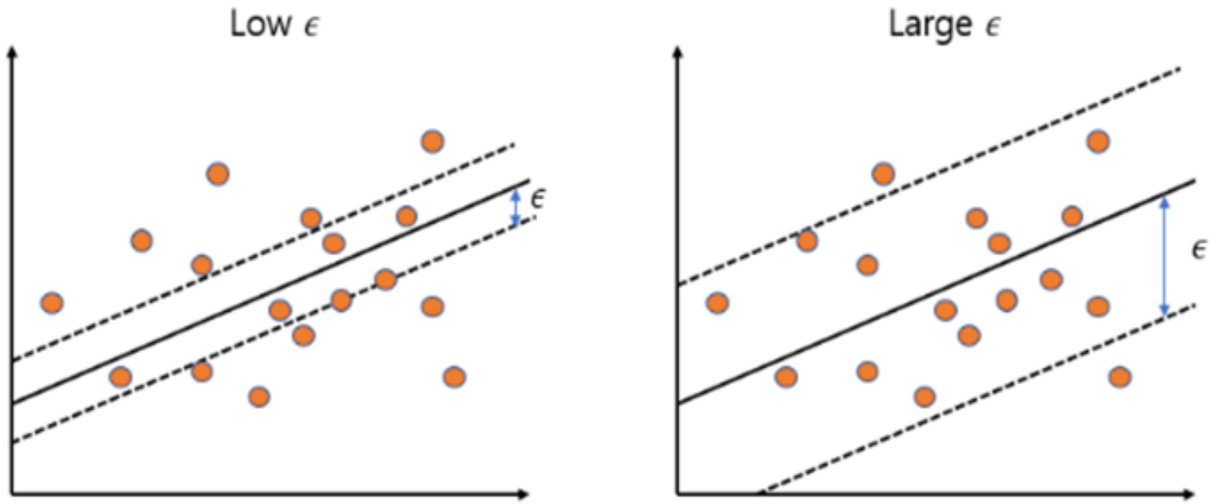
이웃으로 묶는 방법



2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

3) SVR



서포트 벡터 회귀로 SVR은 초평면을 그어 분류 문제로 다루었던 것과 달리 가능한 많은 데이터가 포함된 쪽으로 가운데 초평면을 긋고 그 양 옆으로 데이터가 많이 포함할 수 있게끔 마진 선을 긋는 회귀 모델 초평면을 중심으로 새로운 데이터 수치를 예측하는 모델임.

2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

- 콘텐츠 기반 모델 성능 평가 결과

성능 지표	Knn 베이스 모델	그래디언트 부스팅	Knnregressor	SVR
Precision	0.04347	0.0150	0.0137	0.0151
Recall	0.0082	0.0080	0.0072	0.0094
NDCG	0.00006	0.0164	0.0152	0.0162
Hit	0.00001	0.0016	-	0.0023
RMSE	-	1.0251	1.0551	1.0507

2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

사용자가 최근에 많이 본 장르
모델이 추천한 중복된 영화
모델이 추천한 동일 추천 장르

- **Case Study : USER 1**

기본 베이스 모델은 유저의 history대로 이와 유사한 영화 장르를 보아 코메디 영화들이 유사점이 가장 커서 추천해준거 같음.
그래디언트 부스팅은 영화 장르 정보를 기준으로 회귀 모델을 만들었을 때 액션, 드라마 장르 영화들이 점수가 높을 거라고 예측한 거 같음.
knn은 액션, 드라마 장르 영화 중심으로 그룹화시켜 이 장르의 영화들이 사용자가 더 많이 점수를 줬으니까 사용자에게 추천한 것으로 생각되고, svr은 데이터가 많이 뭉친 대로 회귀 선을 생성해 영화를 추천해줄 때 동일 장르만의 영화들을 더 많이 추천해준 거 같음.

Model : KNN 베이스 모델	
Title	Genres
Back of the Future Part Ⅱ(1989)	Comedy | Sci-Fi
Back of the Future Part Ⅲ(1990)	Comedy | Sci-Fi | Western
Trip to Bountiful, The(1985)	Drama
Father's Day(1997)	Comedy
French Kiss(1995)	Comedy | Romance
Gay Divorce, The(1934)	Comedy | Musical | Romance
Don't Want to Talk about it…….	Drama
Soylent Green(1973)	Sci-Fi | Thriller
Black Cauldron, The(1985)	Animation | Children's
Bridges of Madison Country, The(1995)	Drama | Romance

Model : 그래디언트 부스팅	
Title	Genres
Dead Presidents (1995)	Action | Crime | Drama
Braveheart(1995)	Action| Drama | War
Target (1985)	Action | Drama
Heaven & Earth (1993)	Action | Drama | War
Menace Ⅱ Society (1993)	Action | Crime | Drama
Perfect World, A (1993)	Action | Drama
Program, The(1993).	Action | Drama
Rising Sun (1993)	Action |Drama| Mystery
Romper Stomper (1992)	Action | Drama
Batman (1989)	Action | Adventure | Crime | Drama

Model : Knnregressor	
Title	Genres
Get Shortly (1995)	Action | Crime | Drama
Braveheart(1995)	Action| Drama | War
Target (1985)	Action | Drama
Faster Pussycat! Kill! Kill! (1965)	Action | Comedy| Drama
Heaven & Earth (1993)	Action | Drama| War
Perfect World, A (1993)	Action | Drama
Program, The(1993).	Action | Drama
Rising Sun (1993)	Action |Drama| Mystery
Romper Stomper (1992)	Action | Drama
Last Man Standing (1989)	Action | Adventure | Crime | Drama

Model : SVR	
Title	Genres
BraveHeart (1995)	Action| Drama | War
Heaven & Earth (1993)	Action| Drama | War
Full Metal Jacket (1987)	Action| Drama | War
Boat, The (Das Boot) (1981)	Action| Drama | War
Glory (1989)	Action| Drama | War
G.I. Jane (1997)	Action| Drama | War
Saving Private Ryan (1998).	Action| Drama | War
Thin Red Line, The (1998)	Action| Drama | War
Longest Day, The (1962)	Action| Drama | War
Flying Tigers (1942)	Action| Drama | War

2.4 개인화 추천 시스템

2.4.1 콘텐츠 기반 모델

- **Case Study : USER 1**

장르 분포

Model : KNN 베이스 모델	
Genres	**Count**
Comedy	**5**
Animation	1
Children’s	1
Musical	1
Action	0
Sci-Fi	**3**
Drama	**3**
Western	1
Romance	**3**
Thriller	1
Adventure	0
Crime	0
War	0
Mystery	0

Model : 그래디언트 부스팅	
Genres	**Count**
Comedy	0
Animation	0
Children’s	0
Musical	0
Action	**10**
Sci-Fi	0
Drama	**10**
Western	0
Romance	0
Thriller	0
Adventure	1
Crime	**3**
War	2
Mystery	1

Model : Knnregressor	
Genres	**Count**
Comedy	1
Animation	0
Children’s	0
Musical	0
Action	**10**
Sci-Fi	0
Drama	**10**
Western	0
Romance	0
Thriller	0
Adventure	1
Crime	2
War	2
Mystery	1

Model : SVR	
Genres	**Count**
Comedy	0
Animation	0
Children’s	0
Musical	0
Action	**10**
Sci-Fi	0
Drama	**10**
Western	0
Romance	0
Thriller	0
Adventure	0
Crime	0
War	**10**
Mystery	0

마무리

- 다음과 같은 모델 결과를 봤을 때 불안한 것이 영화 장르 정보 말고도 다른 컬럼 정보를 추가
- 모델을 장르 기반으로 회귀 모델로 구성하면서 동일 장르 영화에 대해서만 점수를 판단하는 경향이 있음.
- 컨텐츠 기반 모델링 방식이 벡터화, 회귀 방식 이외의 다른 방법 탐색

개인화추천시스템

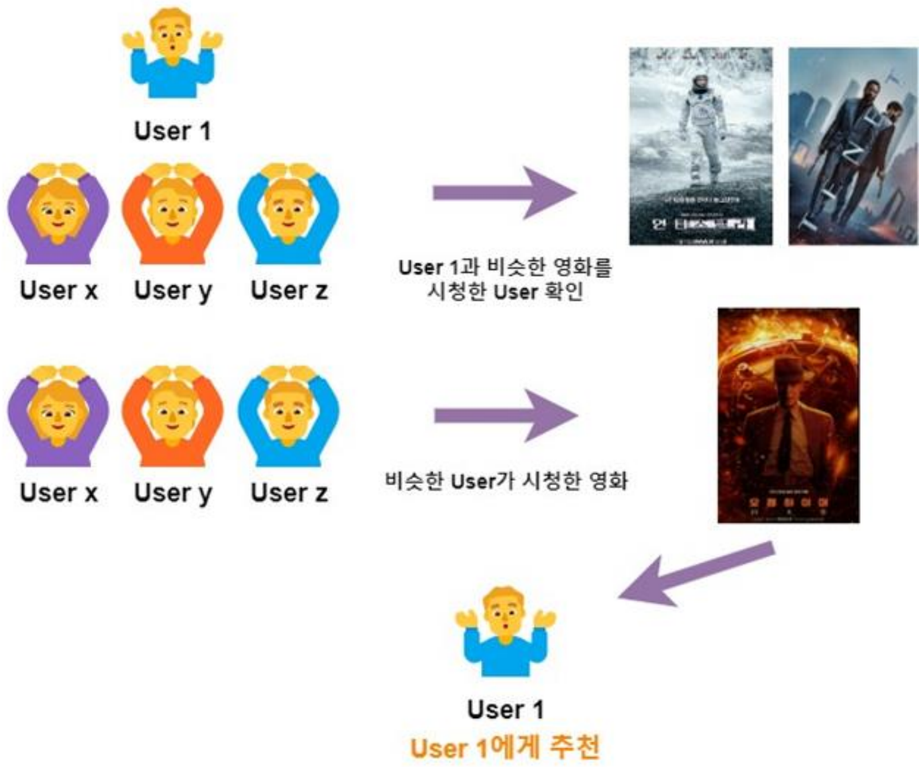
- 협업 필터링 -

발표자 : 권구현

2.4 개인화 추천 시스템

2.4.2 협업 필터링

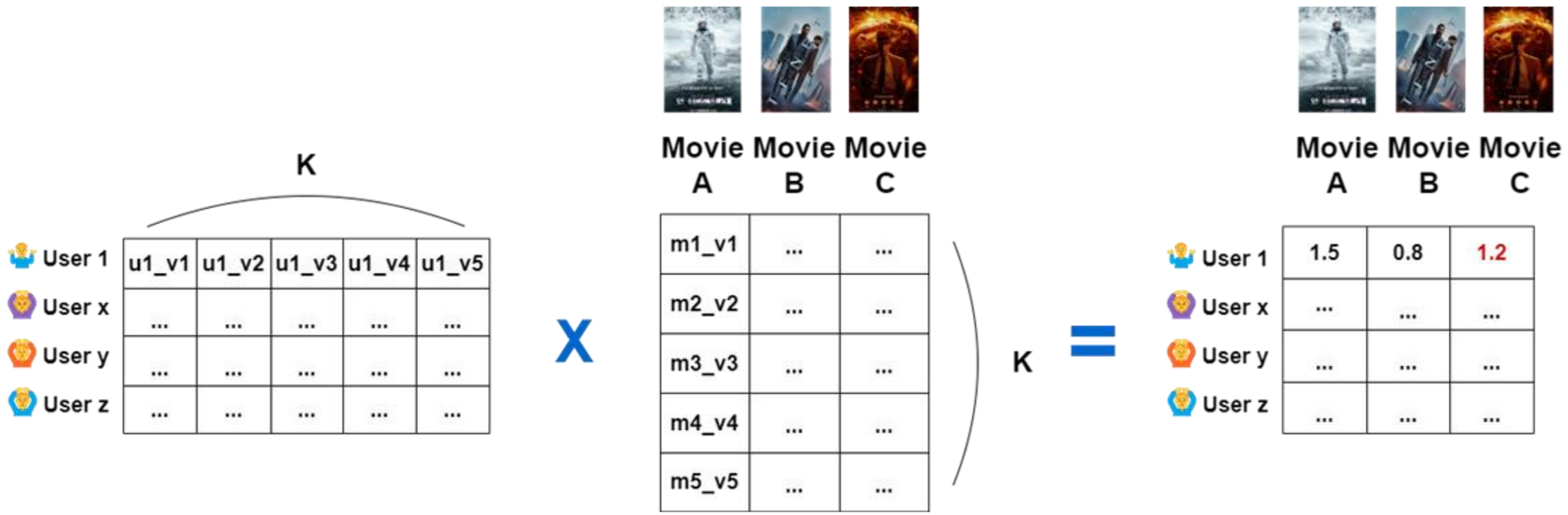
1. **콘텐츠 기반과 다른점** : User 데이터 사용
   - User1과 비슷한 유저가 시청한 영화를 기준으로 추천영화 예측
2. **Matrix Factorization**: 기본적인 협업필터링 모델
   - K개의 유저와 아이템 행렬을 만들어 곱해준 뒤 비어있는 값을 예측하여 사용자에게 추천해주는 방식



2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **Matrix Factorization**



K개의 행렬인 User와 Item 행렬을 곱하여 비어있는 값을 예측

* K의 개수가 커질수록 정확도가 올라가지만, 모델의 복잡도 또한 상승

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **MF 모델 학습 및 예측 결과 확인**

	5 Epoch	20 Epoch	50 Epoch
MSE loss	0.7691	1.0936	1.1502
NDCG @10	0.0331	0.0249	0.0234
HR@10	0.0049	0.0004	0.0003

– 5 Epoch, 20 Epoch, 50 Epoch 성능지표

*Epoch가 올라갈수록 과적합이 심해짐

Title	Genres
Sound of Music, The (1965)	Musical
Schindler's List (1993)	Drama|War
Last Days of Disco, The (1998)	Drama
Cinderella (1950)	Animation|Children's|Musical
Pocahontas (1995)	Animation|Children's|Musical|Romance
Dumbo (1941)	Animation|Children's|Musical
Little Mermaid, The (1989)	Animation|Children's|Comedy|Musical|Romance
Christmas Story, A (1983)	Comedy|Drama
Gone with the Wind (1939)	Drama|Romance|War
Mary Poppins (1964)	Children's|Comedy|Musical

MF모델의 User1에대한 영화 추천리스트

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **MF 모델의 Case Study : USER 1**

- **결과 분석**

MF모델의 User1 추천영화 리스트 분석 결과

User1이 자주 시청했던 영화 위주로 추천 해주었으며

가끔씩 보는 장르도 포함하여 추천하는 것을 확인

- **마무리**

다양한 기법을 통해 성능 고도화 추가 테스트 필요

기본모델 이외의 다른 협업필터링 모델비교 실험 진행

장르 분포

Rank	Genres	Count
1	Drama	4
2	Children's	5
3	Animation	4
4	Comedy	3
5	Musical	6
6	Romance	3
7	Action	0
8	Adventure	0
9	Fantasy	0
10	Sci-Fi	0
11	Thriller	0
12	Crime	0
13	War	2

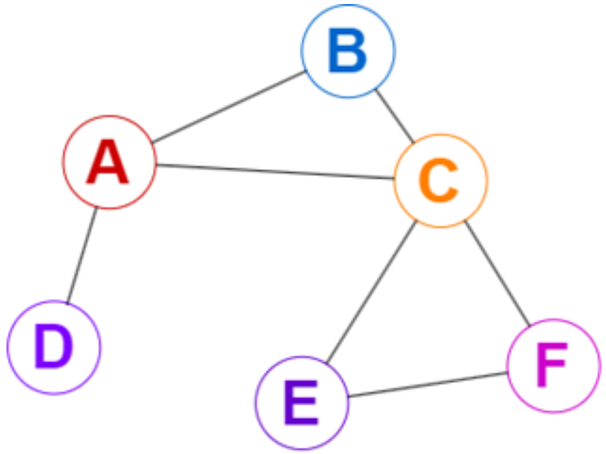
- Rank : User1이 가장 많이 시청한 장르 순위
- Count : MF모델이 예측한 리스트 중 포함된 장르 갯수

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **GNN - Graph Neural Network**

  - User끼리의 간선 그래프를 만들고 인접행렬 생성
  - 인접행렬을 통한 딥러닝 학습 진행
  - positive item, negative item의 임베딩을 생성
  - positive item이 우선순위에 오게하여 User에게 추천영화 출력



–노드 간의 간선 그래프 생성–

	A	B	C	D	E	F
A	0	1	1	1	0	0
B	1	0	1	0	0	0
C	1	1	0	0	1	1
D	1	0	0	0	0	0
E	0	0	1	0	0	1
F	0	0	1	0	0	0

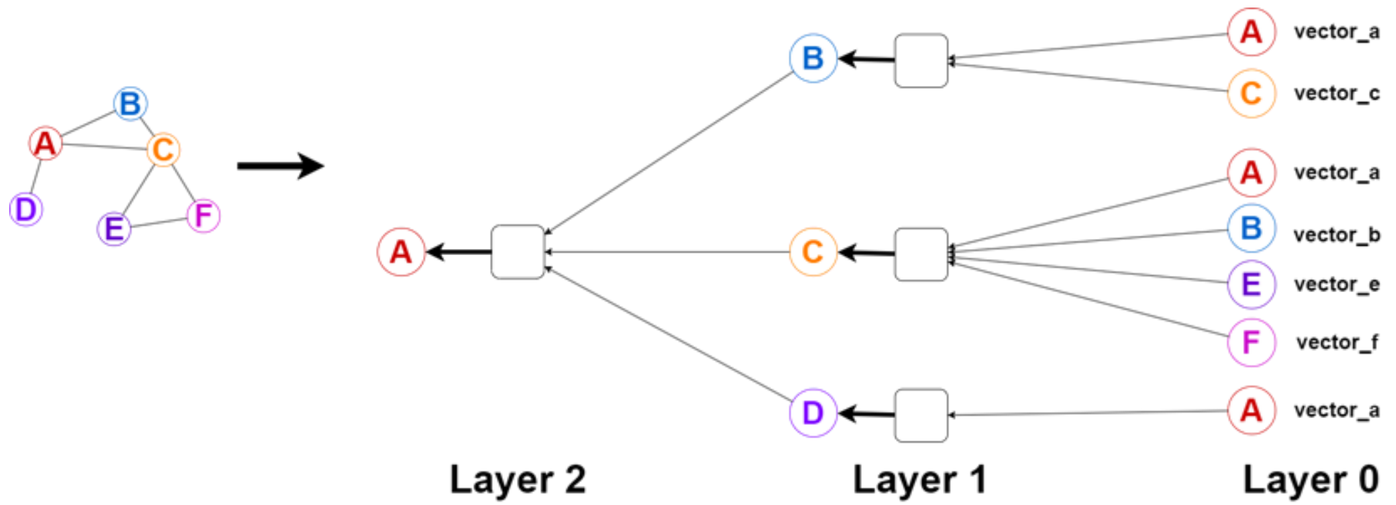
–그래프의 인접 행렬–

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **LightGCN - User, Item 이진관계의 데이터에서 주로 사용**

  - GNN : 다양한 그래프를 분류하며 연산할 수 있는 딥 러닝모델
  - LightGCN : User – Item간의 간선 그래프만 사용하는 이진관계적 데이터에 적합하며 불필요한 연산 간소화



A의 정보를 업데이트하기위해 인접한 B, C, D에게 가중치를 받고

B, C, D 또한 각 인접노드에 하위작업을 통해 가중치를 업데이트 받는 과정을 n_layers 만큼 수행하여 학습

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- 다양한 실험을 적용한 LightGCN 학습

$$\text{BPR loss} = -\text{mean}(\log(\sigma(yui - yuj)))$$

- Basysian Personalized Ranking 수식 -

positive 점수(*yui*)가 negative 점수 (*yuj*) 보다 높아야하는것을 목표로 함

- BPR 방식은 item의 긍정적 점수 스스로 판단 (rating값 미사용)
  *rating값을 적용한 MSE loss 모델과 차이점이 있는지 테스트
- NDCG@10, HR@10성능을 볼때, 기존 BPR loss 모델이 더 우수함

	Base_LGCN	Rating_LGCN
BPR / MSE	0.3331	1.4273
NDCG@10	0.2835	0.2442
HR@10	0.7288	0.0672

Base_LGCN = BPR loss
Rating_LGCN = MSE loss

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **Rating 정보의 차이를 둔 LightGCN 학습 결과**

Title	Genres
American Beauty (1999)	Comedy|Drama
Star Wars: Episode V - The Empire Strikes Back (1980)	Action|Adventure|Drama|Sci-Fi|War
Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Drama|Sci-Fi|War
Jurassic Park (1993)	Action|Adventure|Sci-Fi
Terminator 2: Judgment Day (1991)	Action|Sci-Fi|Thriller
Silence of the Lambs, The (1991)	Drama|Thriller
Matrix, The (1999)	Action|Sci-Fi|Thriller
Men in Black (1997)	Action|Adventure|Comedy|Sci-Fi
Braveheart (1995)	Action|Drama|War
Raiders of the Lost Ark (1981)	Action|Adventure

– BPR loss 사용한 기본 LGCN –

Title	Genres
American Beauty (1999)	Comedy|Drama
Shawshank Redemption, The (1994)	Drama
Raiders of the Lost Ark (1981)	Action|Adventure
Jurassic Park (1993)	Action|Adventure|Sci-Fi
Usual Suspects, The (1995)	Crime|Thriller
Wrong Trousers, The (1993)	Animation|Comedy
Star Wars: Episode VI - Return of the Jedi (1983)	Action|Adventure|Drama|Sci-Fi|War
Casablanca (1942)	Drama|Romance|War
Silence of the Lambs, The (1991)	Drama|Thriller
Braveheart (1995)	Action|Drama|War

– Rating값 추가 후 MSE loss를 사용한 LGCN –

두 모델의 출력결과 중 중복되는 영화가 6개 존재함

2.4 개인화 추천 시스템

2.4.2 협업 필터링

- **LightGCN 모델의 Case Study : USER 1**

• 결과 분석

– MF모델에 비해 사용자의 취향 장르를 예측하는것보다는
전체적인 인지도가 높은 영화를 우선으로 추천하는 듯한 결과를 확인

– 데이터 수가 적은편이기 때문에 모든 그래프가 비슷한 값을 예측

• 마무리

– 기본 User – Item 정보 외에 사용자 프로필, 영화의 메타데이터 등
크고 다양한 데이터를 얻을수록 개인화 추천 정교화 가능성 검증 필요

– 추후 사용할 데이터에따라 Layer, Dropout 등 다양한 기법 적용 필요

장르 분포

Rank	Genres	Count	Count
1	Drama	4	5
2	Children’s	0	0
3	Animation	0	1
4	Comedy	2	2
5	Musical	0	0
6	Romance	1	1
7	Action	7	4
8	Adventure	4	3
9	Fantasy	0	0
10	Sci-Fi	6	2
11	Thriller	3	2
12	Crime	0	1
13	War	3	3

개인화추천시스템

- 하이브리드 -

발표자 : 남자인

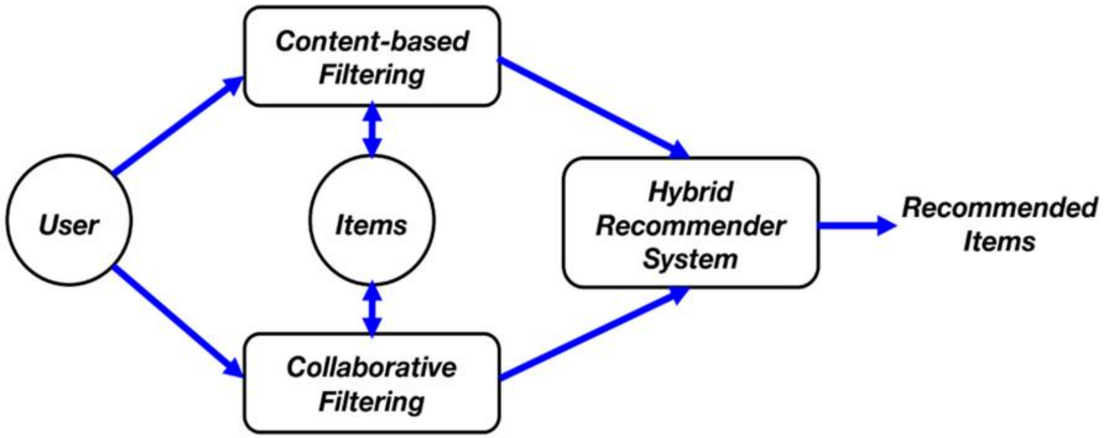
2.4 개인화 추천 시스템

2.4.3 하이브리드

- 하이브리드 모델 개념

  - 각 모델의 장단점을 파악하여 각 방식의 장점을 극대화하면서 단점은 보완하고 다양한 정보를 효과적으로 활용하는 방법

  – 대표적 모델 : 컨텐츠 기반 추천(CBF) + 협업 필터링(CF)



2.4 개인화 추천 시스템

2.4.3 하이브리드

- 하이브리드 모델에 접목하기 위한 협업 필터링 성능 비교

cv=5, n_jobs=4

	KNN_Basic	SVD	NMF	SVD++
RMSE	0.9232	0.8732	0.9174	**0.8622**
MAE	0.7278	0.6857	0.7247	**0.6730**
Fit time	16.31	**3.88**	**4.85**	**220.37**
Test time	44.32	**1.00**	**0.84**	**30.02**

성능은 SVD++ 가 가장 높으나 학습 및 결과를 출력하는 시간이 오래걸림.

그 다음으로 성능이 좋은 SVD 모델과 NMF 모델을 사용

2.4 개인화 추천 시스템

2.4.3 하이브리드

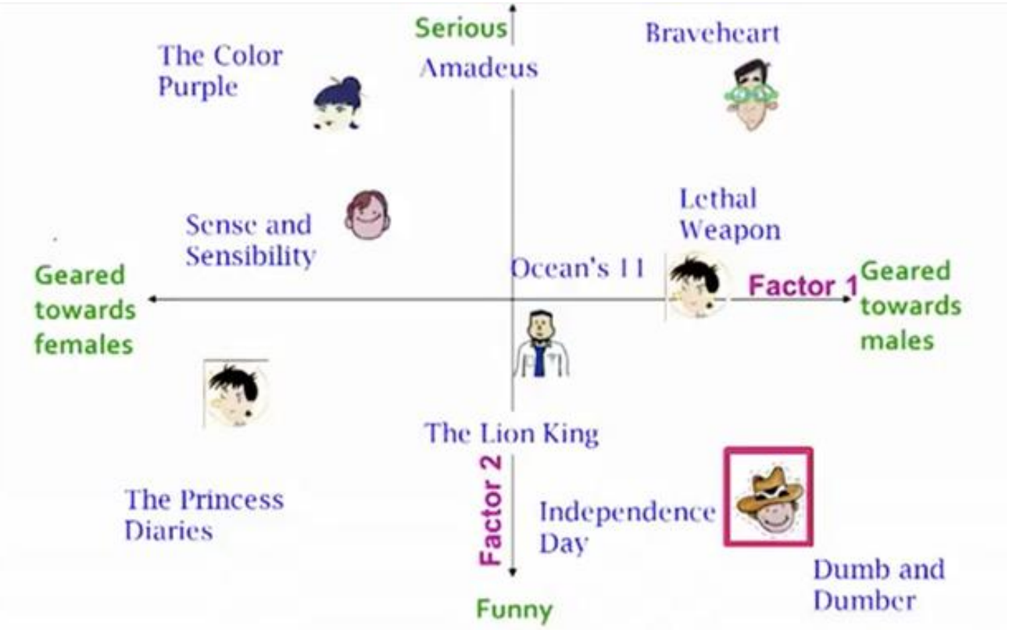
- 사용된 모델 : KNN & MF, SVD, NMF

- **SVD (Singular Value Decomposition)**

  주어진 행렬을 세 개의 행렬의 곱으로 분해하는 기법

- **NMF (Non-Negative Matrix Factorization)**

  양수로만 이루어진 행렬을 여러 개의 양수로만 이루어진 행렬로 분해하는 기법



2.4 개인화 추천 시스템

2.4.3 하이브리드

- 모델 1 : 베이스 코드 기반 모델

- **KNN + MF 모델 : Weighted Ensemble**

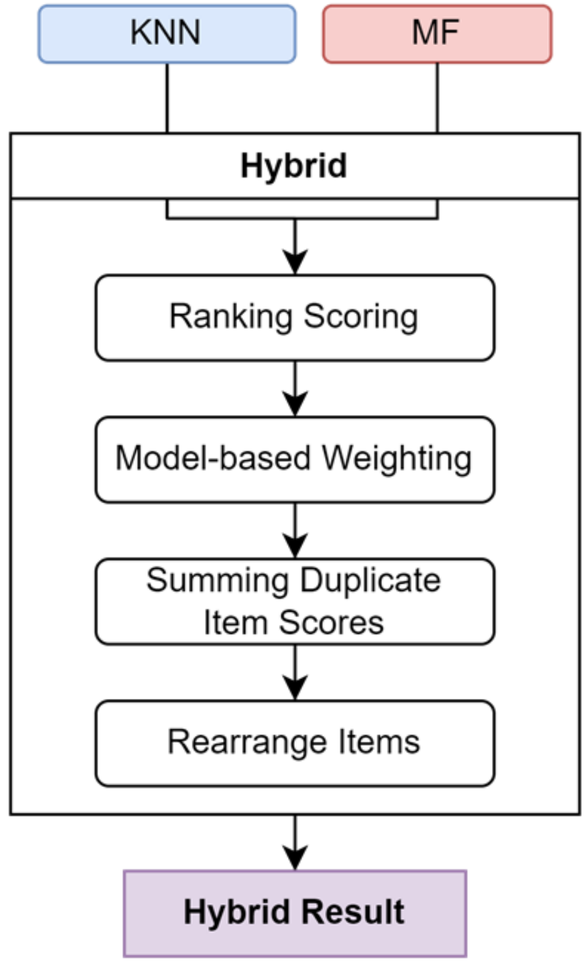
각 결과를 받아

1) 결과별 랭킹에 따른 점수 부여

2) 모델별 가중치에 따른 점수 계산

3) 중복된 아이템의 경우 점수 합산

4) 최종 점수가 부여가 된 아이템을 순서대로 정렬



2.4 개인화 추천 시스템

2.4.3 하이브리드

- 모델 2: SVD 기반 모델

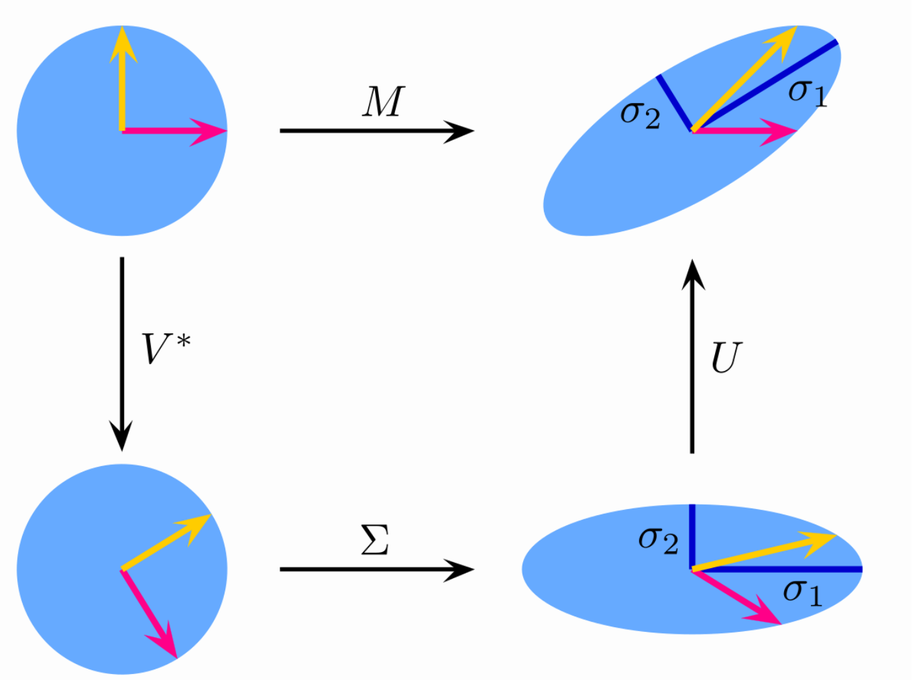
- **SVD+ User-based 모델 :**

1) 각 행과 열이 Movie Id와 User Id로 된 평점 행렬 생성

2) 행렬 분해 수행

3) 특정 사용자에 대한 잠재요인 벡터와 다른 사용자의 벡터 유사도 계산(코사인 유사도)

4) 유사한 사용자가 시청한 영화를 추천



$$M = U \cdot \Sigma \cdot V^*$$

2.4 개인화 추천 시스템

2.4.3 하이브리드

- **모델 3 : NMF 기반 모델**

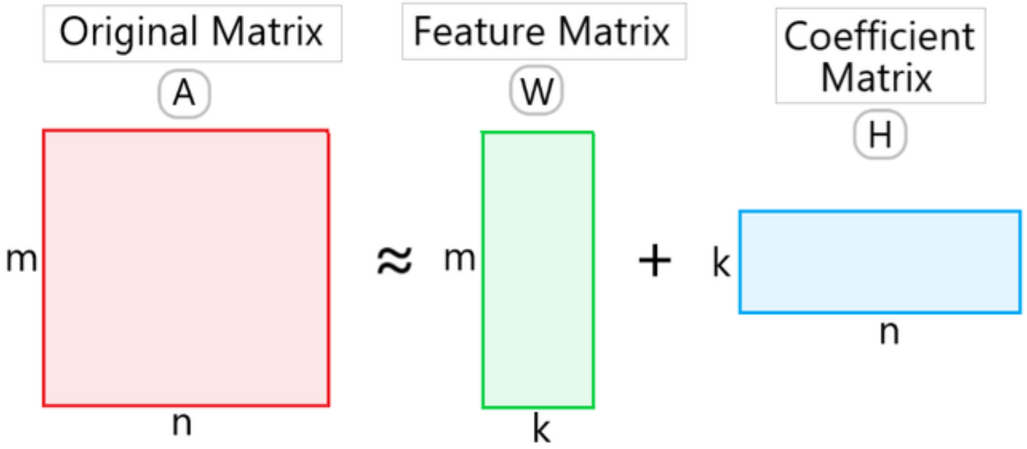
- **NMF + User-based 모델 :**

1) 각 행과 열이 Movie Id와 User Id로 된 평점 행렬 생성

2) 저차원의 비음수 행렬 분해 수행

3) 특정 사용자에 대한 잠재요인 벡터와 다른 사용자의 벡터 유사도 계산(코사인 유사도)

4) 유사한 사용자가 시청한 영화를 추천



2.4 개인화 추천 시스템

2.4.3 하이브리드

- 성능 비교 및 결과 해석

	User Based(Single)	KNN + MF	SVD + User Based	NMF + User Bsed
Precision@10	0.061	0.146	0.158	**0.160**
Recall@10	0.004	0.015	0.010	**0.011**
Hit@10	0	0.002	**0.001**	0.0003
NDCG@10	0.008	0.026	0.026	**0.027**

2.4 개인화 추천 시스템

2.4.3 하이브리드

- **Case Study : USER 1**

Title	Genres
The Fugitibve (1993)	Action|Thriller
Silence of the Lambs, The (1991)	Drama|Thriller
Some Like It Hot (1959)	Comedy|Crime
Winnie the Pooh and the Blustery Day (1968)	Animation|Children's
Godfather: Part II, The (1974)	Action|Crime|Drama
Annie Hall (1977)	Comedy|Romance
Cape Fear (1991)	Thriller
L.A. Confidential (1997)	Crime|Film-Noir|Mystery|Thriller
Jungle Book, The (1967)	Animation|Children's|Comedy
Lady and the Tramp (1955)	Animation|Children's|Comedy

User-Based(Single)

Title	Genres
Sabrina (1995)	Comedy|Romance
GoldenEye (1995)	Action|Adventure|Thriller
American President, The (1995)	Comedy|Drama|Romance
Casino (1995)	Drama|Thriller
Get Shorty (1995)	Action|Comedy|Drama
Powder (1995)	Drama|Sci-Fi
Leaving Las Vegas (1995)	Drama|Romance
Clueless (1995)	Comedy|Romance
To Die For (1995)	Comedy|Drama
Usual Suspects, The (1995)	Crime|Thriller

SVD + User-Based

Title	Genres
Heat (1995)	Action|Crime|Thriller
American President, The (1995)	Comedy|Drama|Romance
Nixon (1995)	Drama
Casino (1995)	Drama|Thriller
Money Train (1995)	Action
Get Shorty (1995)	Action|Comedy|Drama
Copycat (1995)	Crime|Drama|Thriller
Leaving Las Vegas (1995)	Drama|Romance
Dead Man Walking (1995)	Drama
To Die For (1995)	Comedy|Drama

NMF + User-Based

2.4 개인화 추천 시스템

2.4.3 하이브리드

- 하이브리드 모델의 Case Study : USER 1

• 결과 분석

– SVD와 NMF를 적용한 모델은 사용자와 영화간의 잠재적인 특성을 찾아내고 추천하는 경향을 가짐.
– 대부분 드라마나 액션, 스릴러를 추천해주었으나, SVD 모델은 타 모델과 다르게 로맨스를 추천하였음.
– USER 1의 시청 기록을 본다면
두 모델 다 애니메이션, 어린이, 뮤지컬 장르를 추천하지 못했지만,
SVD 모델보다는 NMF가 드라마 장르에서 사용자의 선호도에 근접한 영화를 추천해 줌을 확인할 수 있음.

• 마무리

– 차원을 축소하는 방법인 SVD와 NMF는 다양한 데이터가 부족한 현 데이터 셋에는 적절한 추천이 어려울 수 있음
– 데이터의 정보량 추가 뿐만 아니라 콜드 스타터를 위한 비개인화와 context 를 정보를 반영할 수 있는 딥러닝 모델을 결합한 모델 구축 필요

장르 분포

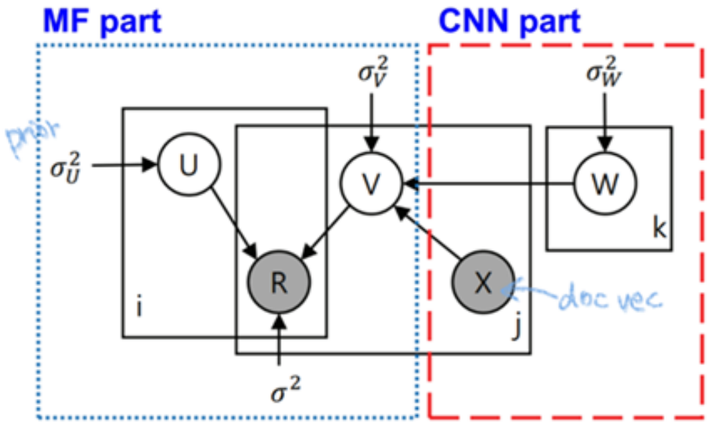
Rank	Genres	Count	Count	Count
1	Drama	2	6	8
2	Children’s	3	0	0
3	Animation	3	0	0
4	Comedy	4	5	2
5	Musical	0	0	0
6	Romance	1	4	2
7	Action	2	2	3
8	Adventure	0	1	0
9	Fantasy	0	0	0
10	Sci-Fi	0	1	0
11	Thriller	4	2	3
12	Crime	0	1	2
13	War	0	0	0

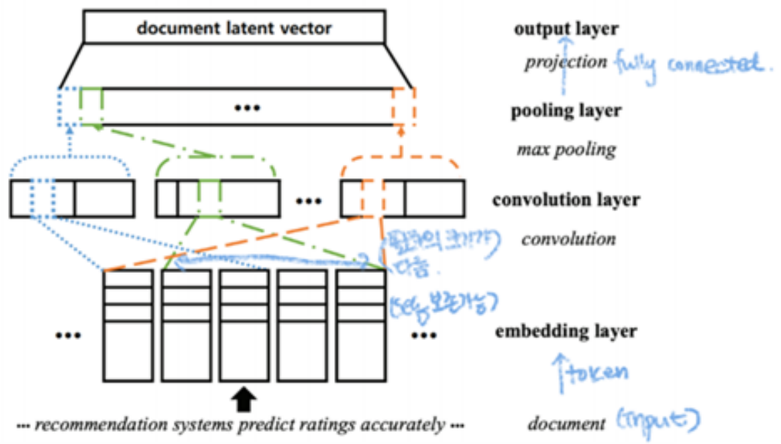
2.4 개인화 추천 시스템

2.4.3 하이브리드

- 하이브리드 모델의 향후 개선 방향

  - 협업 필터링 + 딥러닝 모델 (MF + ANN & MF + CNN)
  - 비개인화 시스템도 결합하여 종합적인 추천 기능 성능향상





개인화추천시스템

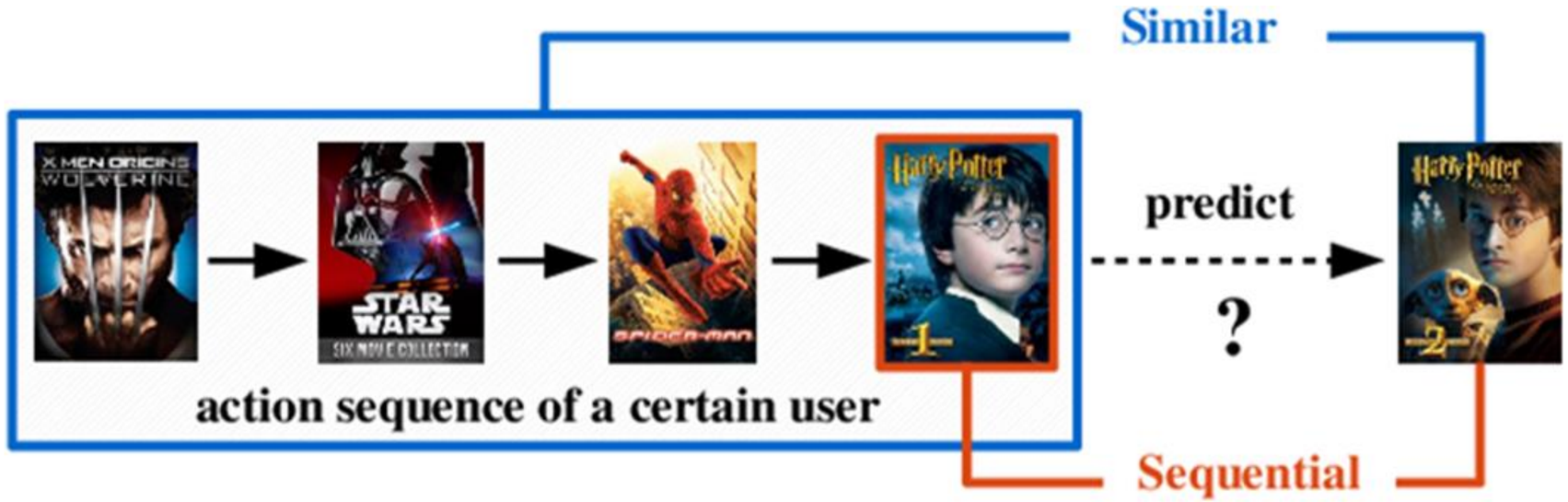
- Sequential Model -

발표자 : 류재영

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **Sequential Model 개념 설명**
  - 사용자가 본 영화를 목록으로 다음 영화를 예측하는 모델



2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **대표적 Sequential System (SASRec, BERT4Rec)**

SASRec

- Self-Attention Mechanism
- Positional Encoding
- Negative Sampling

BERT4Rec

- Bidirectional Self-Attention
- Pretraining
- Textual Understanding

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- 가설 설정

텍스트 이해 능력

플롯 데이터 x

텍스트 리뷰 데이터 x

배우,감독 데이터 x

SASRec이 더 적합(?)

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **RecBole 실행 결과**

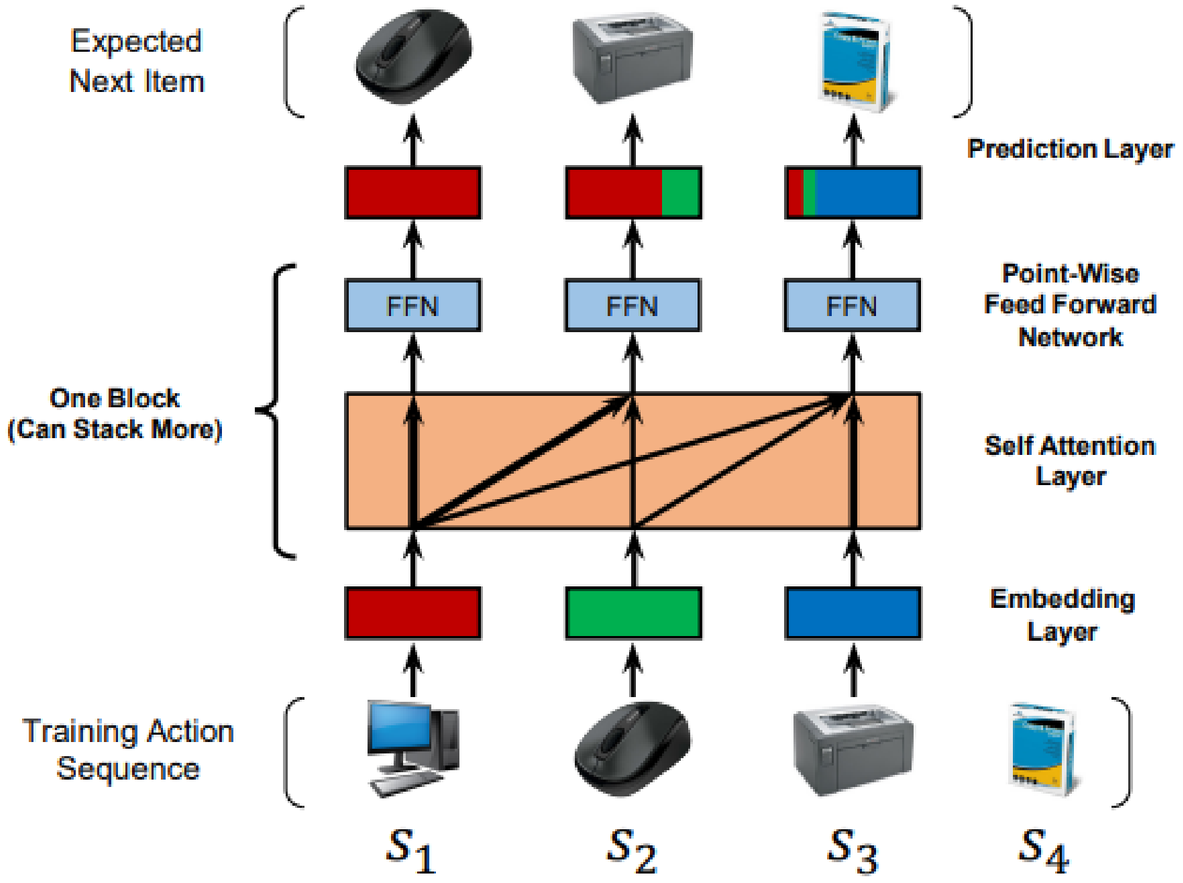
성능 평가 / 모델	SASRec	BERT4Rec
Recall@10	**0.173**	0.0561
Mrr@10	**0.0521**	0.0159
NDCG@10	**0.08**	0.0251
Hit@10	**0.173**	0.0561
Precision@10	**0.0173**	0.0056

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **SASRec 모델 구조 : Architecture**

- Sequential을 고려하기 위해 MC(Markov Chains)와 RNN 이후에 등장한 Transformer는 어텐션 메커니즘인 'self-attention'을 사용하여 의미 있는 패턴을 효율적으로 발견 가능



2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

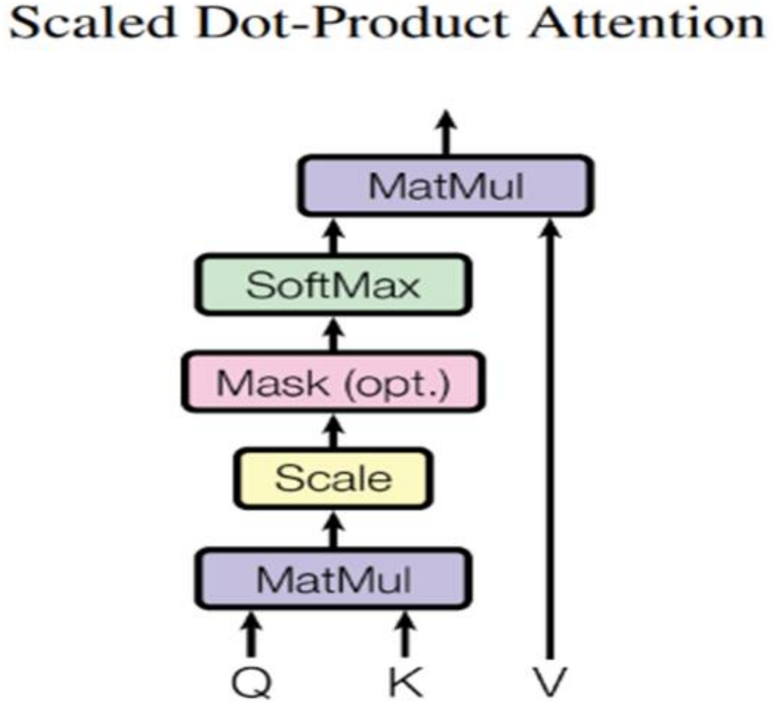
- **SASRec 모델 구조 : Input**
  - Query : 입력 데이터를 나타내는 벡터
  - Key : 입력 데이터와 함께 제공되는 정보를 표현하는 벡터
  - Value : 각 위치의 입력 데이터에 대한 표현을 의미

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **SASRec 모델 구조 : Self-Attention**
  - 유사도 계산
  - 유사도 스케일링 및 정규화
  - 가중 평균 계산

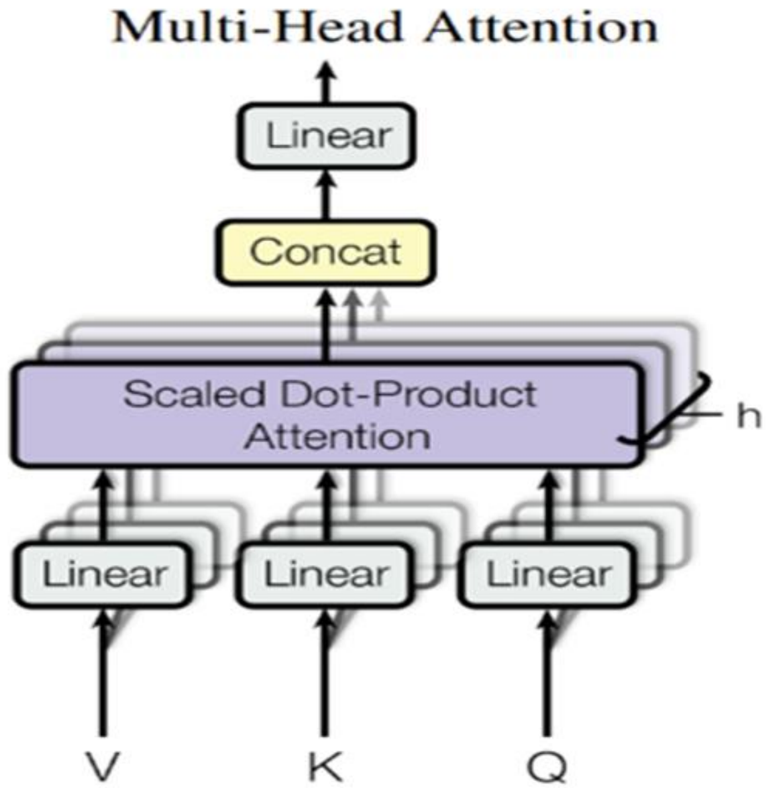
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **SASRec 모델 구조 : Multi Head Attention**
  - 헤드 분리
  - 각 헤드의 어텐션 계산
  - 헤드 결합



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **SASRec 모델 구조 : Input Value**

  - userid : 추천하고자 하는 유저의 아이디
  - seq : 해당 유저가 본 최근 영화 50개(50미만인 유저는 0으로 패딩)
  - time_matrix : 각 요소간의 시간 차이 (패턴과 유사도)
  - item_idx : 예측할 영화 (영화 전체)

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **SASRec 결과**

	Valid_set	Test_set
Epoch	200	
Time	2시간	
NDCG@10	0.5874	0.5818
HR@10	0.8205	0.8083

- **Case Study : USER 1**

Title	Genres	Predicted
Chicken Run (2000)	Animation|Children's|Comedy	6.991471767425537
High Fidelity (2000)	Comedy	6.2796125411987305
Patriot, The (2000)	Action|Drama|War	6.086813449859619
Gladiator (2000)	Action|Drama	5.681070327758789
X-Men (2000)	Action|Sci-Fi	5.39243221282959
Frequency (2000)	Drama|Thriller	5.247025489807129
American Beauty (1999)	Comedy|Drama	4.957281112670898
Cider House Rules, The (1999)	Drama	4.8315300941467285
Fantasia 2000 (1999)	Animation|Children's|Musical	4.648348808288574
Elizabeth (1998)	Drama	4.547706604003906

2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- **Case Study : USER 1 가장 최근에 본 영화 10개**

Title	Genres
Pocahontas (1995)	**Animation**|**Children's**|Musical|Romance
Aladdin (1992)	**Animation**|**Children's**|**Comedy**|Musical
Beauty and the Beast (1991)	**Animation**|**Children's**|Musical
Close Shave, A (1995)	**Animation**|**Comedy**|Thriller
Hunchback of Notre Dame, The (1996)	**Animation**|**Children's**|Musical
Hercules (1997)	Adventure|**Animation**|**Children's**|**Comedy**|Musical
Mulan (1998)	**Animation**|**Children's**
Antz (1998)	**Animation**|**Children's**
Bug's Life, A (1998)	**Animation**|**Children's**|**Comedy**
Tarzan (1999)	**Animation**|**Children's**

- **SASRec 모델의 최상 추천 항목**

	Title	Genres	Predicted
TOP_1	**Chicken Run (2000)**	**Animation|Children's|Comedy**	**6.991471767425537**

* Case Study결과 영화가 잘 추천됨

장르 분포

Genres	Count
Animation	**10**
Children’s	**9**
Musical	**5**
Comedy	**4**
Romance	**1**
Adventure	**1**
Thriller	**1**
Drama	**0**
Action	**0**
Fantasy	**0**
Sci-Fi	**0**
Crime	**0**

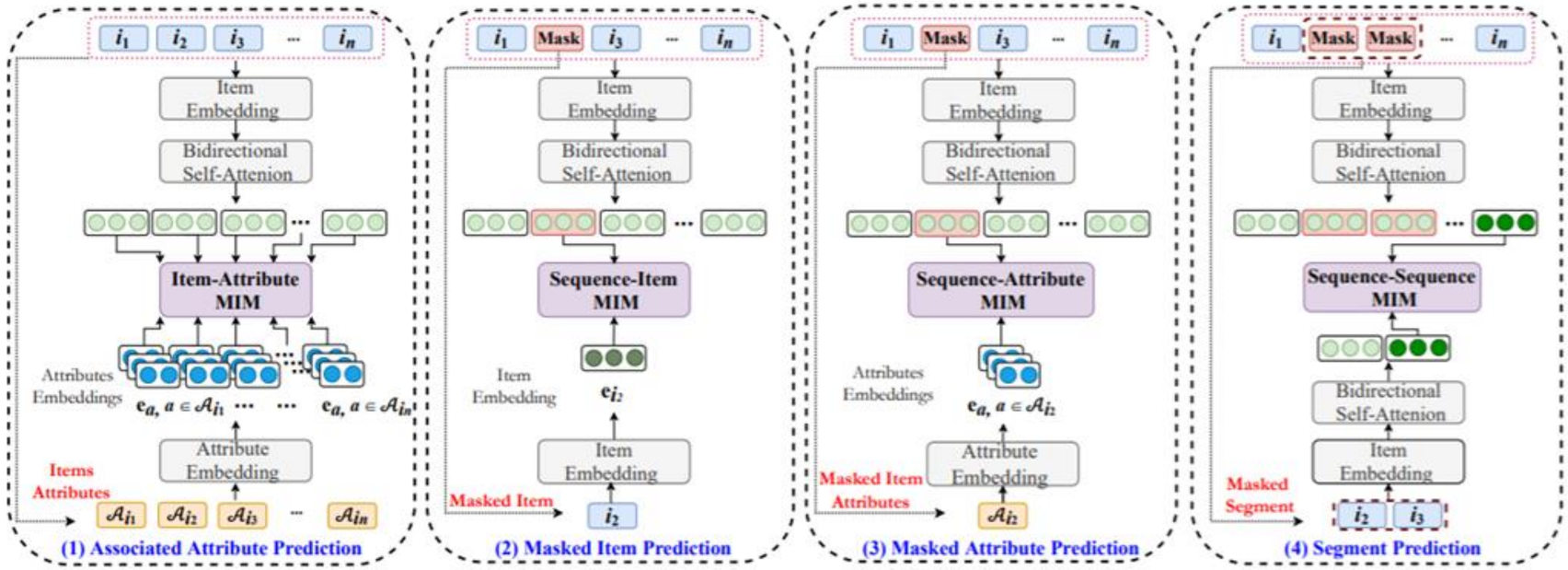
2.4 개인화 추천 시스템

2.4.4 Sequential Recommender System

- 추후 개선 사항

S3Rec 모델은 SASRec 모델에 부가 정보를 사전학습시켜 user-item간 상호작용을 학습, 추론 시키는 모델

사전학습할때는 아래 4개의 방식으로 학습하며, loss는 weight를 다르게 해서 더한 값을 사용



또한 SASRec보다는 최신 모델이고 성능이 높아 보여서 이후 S3Rec 모델의 성능을 확인하고 적용시켜볼 생각입니다.

3. 프로젝트 결과 및 회고

3.1 프로젝트 종합적 결과 설명

3.2 향후 제언

3.3 프로젝트 회고

3.1 프로젝트 최종 결과

프로젝트 종합 결과

- **서비스 제공 방식 제안**

– 컨텐츠 기반, 협업 필터링 기반, 하이브리드, 시퀀셜 모델에서 중복되는 요소 제외 후 N개씩 가져와서 유저에게 10개의 추천리스트를 제공

- **이유**

– 추천알고리즘에서 NDCG@N 의 경우 대중적 지표일 뿐, 절대적 지표가 될 수 없음

즉, 각 모델별 우수한 결과를 선정하여 차츰씩 모델의 비중을 변경하여 AB테스트를 진행하고 서비스에 맞는 모델을 추려나가는 과정 필요

3.2. 향후 제언

종합적 개선 요소

- **성능 고도화**

1) 데이터 정보 추가 (플롯, 배우, 감독, 텍스트 리뷰)

2) Data Augmentation(데이터 증강), 파라미터 조정, Dropout 등 다양한 테스트 작업

3) 추가 모델 탐색

- **시스템 성능 평가**

피드백 시스템 생성(AB test, 설문조사 등 online 평가지표를 통한 시스템 개선)

3.3. 프로젝트 회고

Review

좋았던 점

- 추천 알고리즘에 대한 개념과 내용 이해
- 프로젝트를 진행하며 습득한 팀원과의 긴밀한 협업과정 경험
- 처음 다루는 주제에서 다양한 시도를 통한 문제해결능력 향상
- 기록과 발표작업을 통해 핵심내용 요약 및 공유하는 스킬 습득

아쉬웠던 점

- 다양한 시도를 통해 성능 고도화를 하기 위한 시간이 부족했던 점
- 익숙하지 않은 상황에서 스트레스에 취약했던 점
- 컨디션 및 건강이슈 등 일정에 변수를 주는 요인이 있던 점

감사합니다