



영화 추천 알고리즘

P5 _ 이세나 이정아 오세린 전지우 정윤지



Contents

1. 콘텐츠 기반 필터링
2. 협업 필터링
3. 하이브리드 추천 시스템



1. 콘텐츠 기반 필터링

: 추천 대상 고객 X가 높은 등급을 매긴 아이템과 특징이 유사한 아이템을 고객 X에게 추천

예) 선호하는 영화의 배우, 감독의 다른 영화, 유사 장르의 영화 추천

✓ 진행 단계

- 1) 아이템 프로필 생성
- 2) 사용자 프로필 생성
- 3) 두 행렬 사이의 코사인 유사도 행렬



1. 콘텐츠 기반 필터링

1) 아이템 프로필 생성

아이템별 특징 나타냄

Ex) 아이템 : 영화 특징 : 장르, 배우 등

✓ TF-IDF 방법

TF-IDF로 만들어진 행렬을 토대로 클러스터링 해 대상 아이템과 같은 군집 속에 있는 아이템들을 추천해준다.

즉, 아이템 간의 연관성을 분석해 연관성이 높은 다른 아이템을 찾아서 보여준다.

1. 콘텐츠 기반 필터링

1) 아이템 프로필 생성

✓ TF-IDF 방법

$TF - IDF = TF \times IDF$ (d : 문서, t : 단어, n : 문서의 총 개수)

$TF(d, t)$: 특정 문서 d에서 특정 단어 t의 등장 횟수

$DF(t)$: 특정 단어 t가 등장한 문서의 수

$$IDF(d, t) = \log \frac{n}{(1+DF(t))}$$

※ IDF는 DF의 역수

log : 총 문서의 수 n이 커질 때 IDF의 값이 기하급수적으로 커지는 것 방지

1+DF : 특성 단어가 문서에 등장하지 않을 때 분모가 0이 되는 상황 방지



1. 콘텐츠 기반 필터링

1) 아이템 프로파일 생성

✓ TF-IDF 방법

모든 문서에서 자주 등장하는 단어는 중요도가 낮고
특정 문서에서만 자주 등장하는 단어는 중요도가 높다.

TF-IDF 값이 낮으면 중요도가 낮고 TF-IDF 값이 크면 중요도가 크다.

예) the, a와 같은 불용어는 TF-IDF값이 다른 단어에 비해 낮다.

⇒ TF-IDF방식으로 단어의 가중치를 조정한다.

1. 콘텐츠 기반 필터링

아이템 프로필(TF-IDF) 행렬

	로맨스	스릴러	액션	공상과학	미스터리	코미디	판타지	범죄
A	1.098	0.000	0.000	0.000	0.000	0.000	0.6930	0.000
B	0.000	1.098	0.000	0.000	0.000	0.000	0.6933	0.000
C	0.000	1.098	1.791	0.000	0.000	0.000	0.0000	0.000
D	0.000	0.000	0.000	1.791	0.000	0.000	0.6930	0.000
E	0.000	0.000	0.000	0.000	1.791	0.000	0.0000	1.791
F	1.098	0.000	0.000	0.000	0.000	1.791	0.0000	0.000

1. 콘텐츠 기반 필터링

2) 사용자 프로필 생성

사용자별 아이템 특징에 대한 선호도

TF-IDF 행렬과 사용자의 아이템 선호도 행렬의 내적

사용자 영화 선호도 행렬 (조희수)

	Claudia	Gene	Jack	Lisa	Mick	Toby
A	1	1	0	1	1	0
B	0	1	1	1	1	0
C	1	1	1	1	1	1
D	1	1	1	1	1	1
E	1	1	1	1	1	0
F	1	1	1	1	1	1

영화×사용자



아이템 프로필(TF-IDF) 행렬

	로맨스	스릴러	액션	공상과학	미스터리	코미디	판타지	범죄
A	1.098	0.000	0.000	0.000	0.000	0.000	0.6930	0.000
B	0.000	1.098	0.000	0.000	0.000	0.000	0.6933	0.000
C	0.000	1.098	1.791	0.000	0.000	0.000	0.0000	0.000
D	0.000	0.000	0.000	1.791	0.000	0.000	0.6930	0.000
E	0.000	0.000	0.000	0.000	1.791	0.000	0.0000	1.791
F	1.098	0.000	0.000	0.000	0.000	1.791	0.0000	0.000

영화×장르

1. 콘텐츠 기반 필터링

2) 사용자 프로필 생성

사용자 프로필 행렬

	로맨스	스릴러	액션	공상과학	미스터리	코미디	판타지	범죄
Claudia	2.196	1.098	1.791	1.791	1.791	1.791	1.3860	1.791
Gene	2.196	2.196	1.791	1.791	1.791	1.791	2.0793	1.791
Jack	1.098	2.196	1.791	1.791	1.791	1.791	1.3863	1.791
Lisa	2.196	2.196	1.791	1.791	1.791	1.791	2.0793	1.791
Mick	2.196	2.196	1.791	1.791	1.791	1.791	2.0793	1.791
Toby	1.098	1.098	1.791	1.791	0.000	1.791	0.6930	0.000

사용자×장르

1. 콘텐츠 기반 필터링

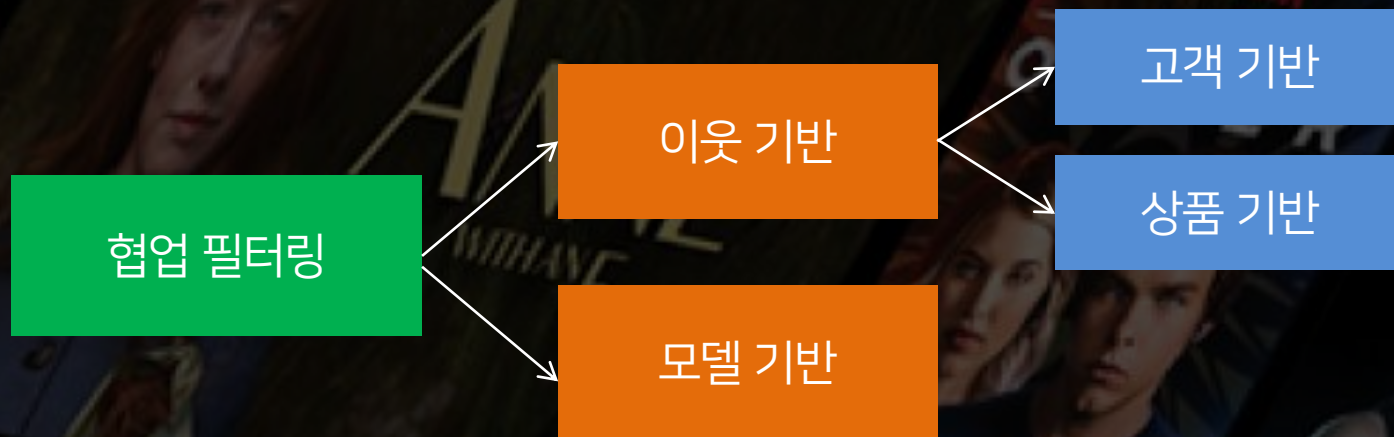
3) 두 행렬 사이의 코사인 유사도 행렬

코사인 유사도가 클수록 사용자가 특정 아이템을 좋아할 가능성이 높음
→ 사용자에게 추천 가능하다고 결론을 내릴 수 있음

	A	B	C	D	E	F
Claudia	0.530210	0.340649	0.428933	0.443163	0.517153	0.546107
Gene	0.541609	0.541629	0.488270	0.441901	0.462382	0.488270
Jack	0.340652	0.530234	0.546097	0.443177	0.517144	0.428925
Lisa	0.541609	0.541629	0.488270	0.441901	0.462382	0.488270
Mick	0.541609	0.541629	0.488270	0.441901	0.462382	0.488270
Toby	0.367031	0.367031	0.593847	0.542856	0.000000	0.593847

2. 협업 필터링

- 가장 널리 사용되며 우수한 기법
- 논문에서 가장 많이 다루는 방식



2. 협업 필터링

1) 이웃 기반

① 고객 기반



2. 협업 필터링

1) 이웃 기반

① 고객 기반

고객 X 상품 선호도 행렬



Step 1. (고객별 상품선호도 기반)
고객간 유사도 측정

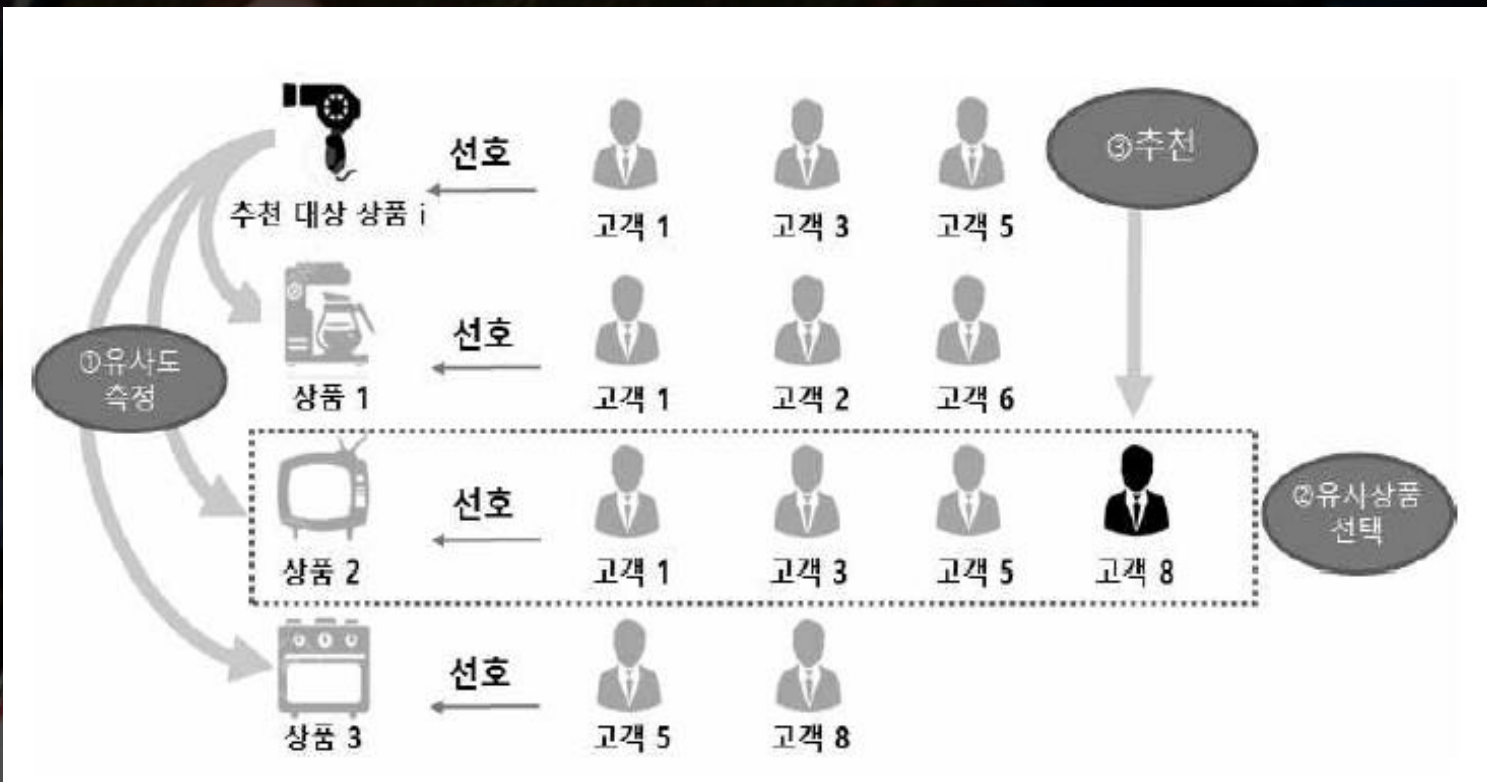
Step 2. 유사도가 가장 가까운 K명의 이웃
고객 선정

Step 3. 선호도 추정 or
추천을 위한 top-N 상품 선정

2. 협업 필터링

1) 이웃 기반

② 상품 기반



2. 협업 필터링

1) 이웃 기반

② 상품 기반

상품 X 고객 선호도 행렬



Step 1. (상품별 고객선호도 기반)
상품간 유사도 측정

Step 2. 유사도가 가장 가까운 K개의 이웃
상품 선정

Step 3. 선호도 추정 or
추천을 위한 top-N 상품 선정

2. 협업 필터링

※ 유사도 측정 방법

고객 간 유사도 혹은 상품 간 유사도 측정

a. 코사인 유사도

-가장 널리 사용됨

-두 벡터(Vector)의 사잇각 코사인 θ 값 이용

$$\text{cosine_sim}(u, v) = \frac{\sum_{i \in I_{uv}} r_{ui} \times r_{vi}}{\sqrt{\sum_{i \in I_{uv}} r_{ui}^2} \times \sqrt{\sum_{i \in I_{uv}} r_{vi}^2}} \quad : \text{사용자간 유사도}$$

$$\text{cosine_sim}(i, j) = \frac{\sum_{u \in U_{ij}} r_{ui} \times r_{uj}}{\sqrt{\sum_{u \in U_{ij}} r_{ui}^2} \times \sqrt{\sum_{u \in U_{ij}} r_{uj}^2}} \quad : \text{아이템간 유사도}$$

I_{uv} : 고객 u와 고객 v가 모두 평가한 상품 집합

r_{ui} : 고객 u의 상품 i에 대한 선호도

r_{vi} : 고객 v의 상품 i에 대한 선호도

U_{ij} : 상품 i와 상품 j를 모두 평가한 고객 집합

r_{uj} : 고객 u의 상품 j에 대한 선호도

2. 협업 필터링

b. 평균 제곱 차이 유사도

-유클리드 공간에서의 거리 제곱에 비례하는 값

$$msd(u, v) = \frac{1}{|I_{uv}|} \sum_{i \in I_{uv}} (r_{ui} - r_{vi})^2$$

$$msd_sim(u, v) = \frac{1}{msd(u, v) + 1}$$

사용자간 유사도

$$msd(i, j) = \frac{1}{|U_{ij}|} \sum_{u \in U_{ij}} (r_{ui} - r_{uj})^2$$

$$msd_sim(i, j) = \frac{1}{msd(i, j) + 1}$$

아이템간 유사도

2. 협업 필터링

c. 피어슨 유사도

두 벡터 간 피어슨 상관계수

$$\text{pearson_sim}(u, v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \mu_u) \times (r_{vi} - \mu_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \mu_u)^2} \times \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \mu_v)^2}} \quad : \text{사용자간 유사도}$$

$$\text{pearson_sim}(i, j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \mu_i) \times (r_{uj} - \mu_j)}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \mu_i)^2} \times \sqrt{\sum_{u \in U_{ij}} (r_{uj} - \mu_j)^2}} \quad : \text{아이템간 유사도}$$

μ_u : 고객 u의 상품 선호도 평균

μ_v : 고객 v의 상품 선호도 평균

μ_i : 상품 i의 고객 선호도 평균

μ_j : 상품 j의 고객 선호도 평균



2. 협업 필터링

※ 선호도 추정 방법

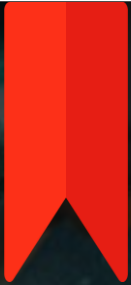
- a. KNNBasic : k명의 이웃 고객/상품의 선호도 점수로 단순평균을 추정값으로 사용
- b. KNNwithMeans : 선호도 점수들의 평균값을 기준으로 가중평균값을 추정값으로 사용
- c. KNNwithZscore : 고객별로 Z-표준화 선호도 점수를 계산해 반영한 값을 추정값으로 사용
- d. KNNBaseline : 선호도 점수들의 베이스라인 모델의 값을 기준으로 한 가중평균값을 추정값으로 사용



2. 협업 필터링

2) 모델 기반

- 고객-상품 간의 선호도 정보를 선호도 추정 모델 트레이닝에 사용
- 잠재요인 모델(Latent Factor Model) : 고객-상품 간 상호작용을 그들의 잠재적 특성(Latent characteristics)을 나타내는 요인(Factor)으로 모델링
- 고객/상품의 특성 벡터의 크기는 수천, 수백만에 달할 수 있으므로 이를 몇 개의 잠재 요인 벡터로 간략화
- 향후 특정 고객에 대해 새로운 상품을 추천하기 위한 선호도 추정에 사용



2. 협업 필터링

2) 모델 기반

행렬 분해(matrix factorization) 모델

- 고객 벡터 또는 상품 벡터를 PCA로 축소하거나 SVD로 분해
- 고객에 대한 잠재요인(Latent factor)과 상품에 대한 잠재요인을 추출

$$R \approx PQ^T$$

$R \in R^{m \times n}$: m명 고객과 n개 상품의 선호도 행렬

$P \in R^{m \times k}$: m명 고객과 k개 요인의 관계 행렬

$Q \in R^{n \times k}$: n개 상품과 k개 요인의 관계 행렬

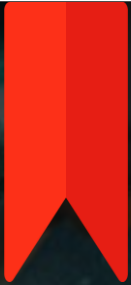
2. 협업 필터링

2) 모델 기반

	M1	M2	M3	M4	M5
 Comedy	3	1	1	3	1
 Action	1	2	4	1	3

	 Comedy	 Action
 A	✓	✗
 B	✗	✓
 C	✓	✗
 D	✓	✓

	M1	M2	M3	M4	M5
 A	3	1	1	3	1
 B	1	2	4	1	3
 C	3	1	1	3	1
 D	4	3	5	4	4



2. 협업 필터링

2) 모델 기반

SVD(Singular Value Decomposition) 모델

- 행렬분해 방법 중 하나
- Netflix 경진대회에서 Simon Funk 사용하면서 유명해짐

$$R = U\Sigma V^T$$

R = 선호도 행렬(m명 고객 x n개 상품)

U = 사용자 행렬 Σ = 특이값 행렬 V=아이템 행렬

2. 협업 필터링

2) 모델 기반

SVD(Singular Value Decomposition) 모델

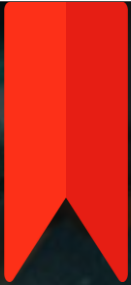
- 각 행렬에서 k개만 추출

$$\hat{U} \hat{\Sigma} \hat{V}^T = \hat{R} \approx R$$

$\hat{\Sigma}$: Σ 의 대각 성분 중 가장 큰 k개의 특잇값을 추출한 k×k 크기의 대각 행렬

\hat{U} : U 행렬에서 가장 큰 k개의 특잇값에 대응하는 k개의 성분만을 남긴 m×k 크기의 행렬

\hat{V} : V 행렬에서 가장 큰 k개의 특잇값에 대응하는 k개의 성분만을 남긴 k×n 크기의 행렬



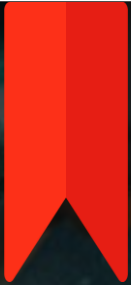
2. 협업 필터링

2) 모델 기반

SVD(Singular Value Decomposition) 모델

$\hat{U}\sqrt{\hat{\Sigma}}^T$ = 사용자와 요인 ($m \times k$)

$\sqrt{\hat{\Sigma}}\hat{V}^T$ = 아이템과 요인 ($k \times n$)



2. 협업 필터링

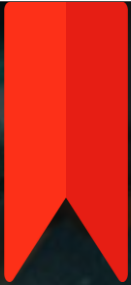
2) 모델 기반

SVD(Singular Value Decomposition) 모델

선호도 행렬, 평점 행렬 : 빈 원소가 많으므로 SVD를 바로 적용하기 어려움

→ 기존 데이터 행렬과 잠재 변수 행렬의 내적곱 행렬(PQ^T) 간의
제공 오차를 최소화하는 방식으로 결측치를 채움

→이 때, 과적합 방지를 위해 SGD(Stochastic Gradient Descent)와
ALS(Alternating Least Squares) 같은 정규화 방법 이용



2. 협업 필터링

2) 모델 기반

이외에도 SVD++ 모델, Non-negative matrix factorization 모델, 베이스라인 모델, slope one 모델, co clustering 모델 등

→ 현존하는 가장 우수한 추천 알고리즘은 SVD++

2. 협업 필터링

모델 평가

- 1) 사용자가 실제 평가한 아이템에 대한 예상 평점 예측
→ 실제 평점과 차이 계산

⇒ 평가 지표가 작을수록 모형이 정확

$$a. MAD = \frac{\sum_{i=1}^n \sum_{j=1}^m (|r_{i,j} - p_{i,j}| \text{ if } r_{i,j} \neq \text{null}, 0 \text{ otherwise})}{k}$$

$$b. MSE = \frac{\sum_{i=1}^n \sum_{j=1}^m ((r_{i,j} - p_{i,j})^2 \text{ if } r_{i,j} \neq \text{null}, 0 \text{ otherwise})}{k}$$

$$c. RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^m ((r_{i,j} - p_{i,j})^2 \text{ if } r_{i,j} \neq \text{null}, 0 \text{ otherwise})}{k}}$$

k : 정확도 계산에 포함된 총 아이템 수
n : 검증용 데이터에 있는 사용자 수
m : 검증용 데이터에 있는 아이템 수
 $r_{i,j}$: 사용자 i의 아이템 j에 대한 실제 평점
 $p_{i,j}$: 사용자 i의 아이템 j에 대한 예상 평점

2. 협업 필터링

모델 평가

2) 추천한 아이템과 실제 선택된 아이템을 비교

a. 정확도(Precision, Accuracy) = $\frac{\text{사용자가 실제 선택한 아이템 수}}{\text{전체 추천된 아이템 수}}$

b. 재현율(Recall) = $\frac{\text{맞는 추천 아이템 수}}{\text{사용자가 선택한 전체 아이템 수}}$

c. F 측정값(F measure) = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$



3. 하이브리드 추천 시스템

다양한 연구들이 이뤄지고 있음

1) 콘텐츠 기반 필터링+협업 필터링

-Combining Filtering 기법 :

콘텐츠 기반 필터링과 협업 필터링의 알고리즘을 모두 적용하고 가중평균(weighted average)을 구하는 방법

-Collaborative via Content 기법 :

평점 데이터와 아이템 프로필을 조합해 사용자 프로필을 만들어 추천하는 방법

2) 사용자 기반+아이템 기반 협업 필터링

-사용자 기반 협업 필터링을 이용해 유사한 선호도를 가진 이웃집단을 형성 & 아이템 간의 선호도를 토대로 이웃집단을 형성

→ 가장 적합한 상품을 고객에게 추천한다.



추후 계획

컨텐츠 기반 필터링(장르 외에도 배우, 감독을 아이템으로 설정해 결합할 예정),
협업 필터링의 고객 기반, 상품 기반, 모델 기반 필터링 시스템에 대한
코딩 작업을 진행할 예정!



참고자료

1. 콘텐츠 기반 필터링

<https://wikidocs.net/31698>

<https://movefast.tistory.com/238>

<https://rstatistics.tistory.com/28#-->

2. 협업 필터링

<https://data-science-hi.tistory.com/82?category=1077184>

<https://jeongchul.tistory.com/553>

벡터공간모델을 활용한 상품추천 알고리즘에 관한 실증연구(정성원)

행렬 분해 및 잠재 변수를 활용한 협업 필터링 기반 텍스트 콘텐츠 추천(윤주식)



Q & A



THANK YOU!