DATA CREATOR CAMP

2024 데이터 크리에이터 캠프

대학부 실습영상

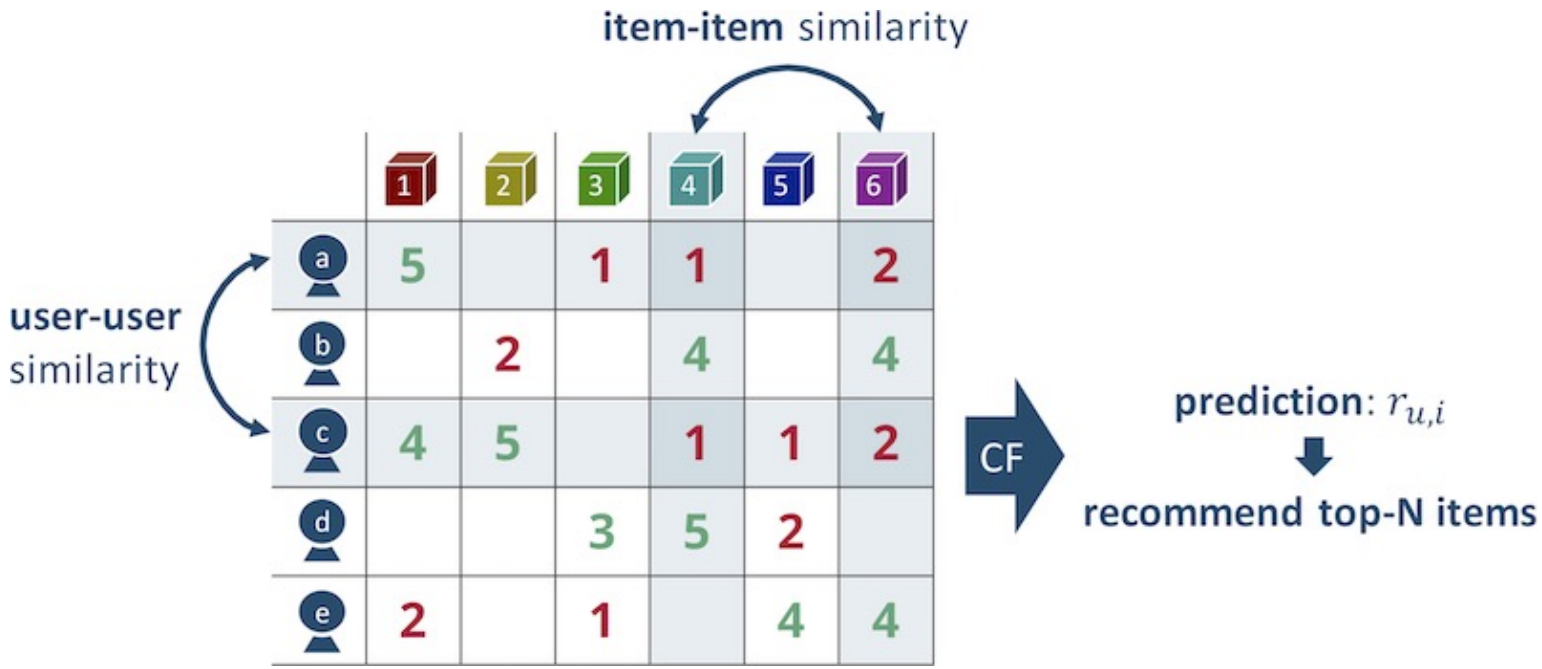
7강. Collaborative Filtering

목차

1. 협업 필터링의 유형

❖ 협업 필터링의 정의

- 사용자와 아이템 간의 상호작용 데이터를 기반으로 유사한 사용자나 아이템을 찾아서 새로운 추천을 생성하는 방법
- 사용자의 과거 행동 데이터 또는 유사한 사용자 그룹의 데이터를 활용하여 개인화된 추천을 제공.



https://blog.kakaocdn.net/dn/rw0nf/btrhTgs8Soh/0dNBT6BywKXy4I1KKxQ5S0/img.png

1. 협업 필터링의 유형

* 6강 참고

- User-based collaborative filtering
  - 이전에 시청한 영화들 중 비슷하게 평가한 사용자를 통해서 타겟 사용자의 타겟 아이템에 대한 점수를 예측함
  - 사용자간의 유사성 평가는 row를 기준으로 구함
- Item-based collaborative filtering
  - 타겟 사용자가 이전에 시청한 영화들과 비슷한 영화를 찾아, 타겟 사용자의 평가를 바탕으로 타겟 아이템에 대한 점수를 예측함.
  - 아이템간 유사성 평가는 column를 기준으로 구함

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?
사용자3	1	5	3	3	5	1
사용자4	2	?	2	1	4	1

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?
사용자3	1	5	4	4	5	1
사용자4	2	?	4	5	4	1

2. 협업 필터링 과정

① 데이터 준비: 사용자-아이템 평가 데이터를 바탕으로 유틸리티 행렬을 생성.

② 유사도 계산: 사용자 간 또는 아이템 간 유사도를 계산하여 비슷한 사용자나 아이템을 찾음.

③ 평점 예측: 유사도를 바탕으로 아직 평가되지 않은 아이템에 대한 예상 평점을 계산.

④ 추천 제공: 예측된 평점이 높은 아이템을 추천.

⑤ 피드백과 반복: 사용자의 피드백을 반영하여 추천 정확도를 지속적으로 향상.

2. 협업 필터링 과정

① 데이터 준비

- 사용자-아이템 상호작용 데이터를 준비
- 사용자-아이템 상호작용 데이터 : 사용자가 특정 아이템에 대해 어떻게 평가했는지를 기록한 평가 행렬(utility matrix)
  - 행(row)은 사용자, 열(column)은 아이템을 나타냄
  - 각 셀에는 사용자가 해당 아이템에 대해 부여한 평가 점수(예: 별점)가 들어감
  - 비어 있는 값(누락된 값)이 있을 수 있으며, 이는 사용자가 해당 아이템에 대해 평가하지 않았음을 의미함.

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?
사용자3	1	5	3	3	5	1
사용자4	2	?	2	1	4	1

2. 협업 필터링 과정

② 유사도 계산

- 사용자 또는 아이템 간의 유사도(similarity)를 계산
- 사용자 기반 협업 필터링 : 비슷한 성향을 가진 사용자(유사한 평가 패턴을 보이는 사용자)를 찾고, 사용자의 평가 데이터의 유사성을 기반
- 아이템 기반 협업 필터링: 특정 사용자가 평가한 아이템과 유사한 아이템을 찾아 추천합니다. 아이템 간의 유사도는 사용자가 해당 아이템에 부여한 평가 점수의 유사성을 기반

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?
사용자3	1	5	3	3	5	1
사용자4	2	?	2	1	4	1

2. 협업 필터링 과정

③ 평점 예측

- 유사도 계산을 바탕으로 사용자가 아직 평가하지 않은 아이템의 평점을 예측
- 사용자 기반 협업 필터링: 사용자 간 유사도에 기반하여, 타겟 사용자의 유사한 사용자가 특정 아이템에 부여한 평가 점수를 사용해 타겟 사용자의 평점을 예측
- 아이템 기반 협업 필터링: 사용자가 평가한 유사한 아이템의 평가 점수를 바탕으로 새로운 아이템에 대한 사용자의 평점을 예측

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	5
사용자3	1	5	3	3	5	1
사용자4	2	5	2	1	4	1

2. 협업 필터링 과정

④ 추천 제공

- 예측된 평점을 바탕으로 추천 리스트를 생성하고, 사용자가 관심을 가질 만한 아이템을 추천
- 예측된 평점이 높은 순서대로 아이템을 정렬한 후, 사용자가 아직 평가하지 않은 상위 아이템을 추천 리스트에 포함시킴.

⑤ 피드백과 반복

- 추천 시스템은 사용자의 피드백을 통해 점점 더 나은 추천을 제공함.
- 사용자가 추천된 아이템을 평가하거나 소비할 때, 이 데이터는 평가 행렬에 추가되며, 이를 기반으로 다음 추천이 이루어짐.

2. 협업 필터링 과정

❖ User-Based Neighborhood Models

▪ 사용자 기반의 아이템 평점 예측

사용자 u와 사용자 v의 유사정도

사용자 u와 유사한 사용자들

사용자 v의 아이템 j에 대한 평점

사용자 u의 아이템 j에 대한 예측 평점

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} \mathrm{Sim}(u,v) \cdot s_{vj}}{\sum_{v \in P_u(j)} |\mathrm{Sim}(u,v)|}$$

사용자 u의 평균 평점

: 기본적으로 사용자가 보통 얼마나 높거나 낮게 평가하는지를 반영

사용자 v가 아이템 j에 대해 평가한 평균 중심화된 평점

$$= \mu_u + \frac{\sum_{v \in P_u(j)} \mathrm{Sim}(u,v) \cdot (r_{vj} - \mu_v)}{\sum_{v \in P_u(j)} |\mathrm{Sim}(u,v)|} \qquad (2.4)$$

2. 협업 필터링 과정

❖ User-Based Neighborhood Models

▪ 사용자 기반의 아이템 평점 예측

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장	sim
사용자1	5	1	3	2	2	5	0.79
사용자2	5	2	2	2	2	?	0.75
사용자3	1	5	4	4	5	1	0.97
사용자4	2	5	4	5	4	1	

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} \mathrm{Sim}(u,v) \cdot s_{vj}}{\sum_{v \in P_u(j)} |\mathrm{Sim}(u,v)|} = \mu_u + \frac{\sum_{v \in P_u(j)} \mathrm{Sim}(u,v) \cdot (r_{vj} - \mu_v)}{\sum_{v \in P_u(j)} |\mathrm{Sim}(u,v)|} \qquad (2.4)$$

- Pu(j) = 사용자 4와 유사도가 0.8이상인 사용자 = {사용자3}
- 사용자 4의 검은 사제들에 대한 예측 평점

  = 3.2 + (0.97 * (5 – 3.2)) / 0.97 = 5

3. 유사도 함수

❖ 유사도 함수 종류

- Sim(u,v)를 구하는 데에 사용되는 유사도 함수는 다양합니다.

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} \mathrm{Sim}(u, v) \cdot s_{vj}}{\sum_{v \in P_u(j)} |\mathrm{Sim}(u, v)|}$$

3. 유사도 함수

❖ MSD(mean squared difference)

- 두 사용자 또는 두 아이템이 공통적으로 평가한 항목들의 평가값 차이를 제곱한 후 평균을 계산
- 낮은 MSD 값일수록 두 사용자가 유사한 평가 패턴을 보인다는 것을 의미합니다.

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

$|I_{uv}|$: 공통으로 평가한 아이템의 수

$r_{u,i}$: 사용자 u가 아이템 i에 부여한 평점

3. 유사도 함수

❖ MSD(mean squared difference)

- 두 사용자 또는 두 아이템이 공통적으로 평가한 항목들의 평가값 차이를 제곱한 후 평균을 계산
- 낮은 MSD 값일수록 두 사용자가 유사한 평가 패턴을 보인다는 것을 의미합니다.

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?

$$\mathrm{MSD}(A,B) = \frac{0+1+1+0+0}{5} = 0.4$$

3. 유사도 함수

❖ Cosine Similarity (코사인 유사도)

- 두 벡터 간의 코사인 각도를 기반으로 유사도를 계산
- 두 벡터가 얼마나 유사한지를 측정하는 방법으로, 0에서 1 사이의 값을 가집니다.
- 1에 가까울수록 유사함.

$$\text{Cosine Similarity}(u, v) = \frac{\sum_{i \in I_{uv}} r_{u,i} \cdot r_{v,i}}{\sqrt{\sum_{i \in I_u} r_{u,i}^2} \cdot \sqrt{\sum_{i \in I_v} r_{v,i}^2}}$$

3. 유사도 함수

$$\text{Cosine Similarity}(u, v) = \frac{\sum_{i \in I_{uv}} r_{u,i} \cdot r_{v,i}}{\sqrt{\sum_{i \in I_u} r_{u,i}^2} \cdot \sqrt{\sum_{i \in I_v} r_{v,i}^2}}$$

❖ Cosine Similarity (코사인 유사도)

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?

- 사용자 A의 평가 벡터: [5, 1, 3, 2, 2, 5]
- 사용자 B의 평가 벡터: [5, 2, 2, 2, 2, ?] (결측값은 0으로 처리)
- 두 벡터 간의 내적:

$$\sum_{i \in I_{AB}} r_{A,i} \cdot r_{B,i} = (5 \times 5) + (1 \times 2) + (3 \times 2) + (2 \times 2) + (2 \times 2)$$

- 사용자 A의 벡터 길이:

$$\sqrt{\sum_{i \in I_A} r_{A,i}^2} = \sqrt{5^2 + 1^2 + 3^2 + 2^2 + 2^2 + 5^2} = \sqrt{25 + 1 + 9 + 4 +$$

- 사용자 B의 벡터 길이:

$$\sqrt{\sum_{i \in I_B} r_{B,i}^2} = \sqrt{5^2 + 2^2 + 2^2 + 2^2 + 2^2 + 0^2} = \sqrt{25 + 4 + 4 + 4 -$$

- 코사인 유사도:

$$\text{Cosine Similarity}(A, B) = \frac{41}{8.25 \times 6.40} = 0.776$$

3. 유사도 함수

❖ Pearson Correlation (피어슨 상관계수)

- 두 사용자 간의 평가 값이 선형적으로 얼마나 상관관계가 있는지를 측정
- 두 사용자 간의 평가 패턴이 얼마나 일관되게 움직이는지를 계산
- 피어슨 상관계수는 -1에서 1 사이의 값을 가지며, 1에 가까울수록 두 사용자 간의 선형 상관관계가 높음

사용자 u의 아이템 i에 대한 평점

사용자 u의 평균평점

$$\text{Pearson Correlation}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \mu_u) \cdot (r_{v,i} - \mu_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \mu_u)^2} \cdot \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \mu_v)}}$$

사용자 u의 분산

사용자 v의 분산

3. 유사도 함수

❖ Pearson Correlation (피어슨 상관계수)

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?

$$\text{Pearson Correlation}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \mu_u) \cdot (r_{v,i} - \mu_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \mu_u)^2} \cdot \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \mu_v)}}$$

$$\text{Pearson Correlation}(A,B) = \frac{7.2}{3.16 \times 2.68} = 0.885$$

3. 유사도 함수

❖ Jaccard Similarity (자카드 유사도)

- 두 집합 간의 교집합 크기를 합집합 크기로 나누어 유사도를 계산하는 방법
- 평가값보다는 집합의 유사도를 계산하는 데 사용
- 자카드 유사도는 0에서 1 사이의 값을 가지며, 1에 가까울수록 두 집합이 유사하다는 의미함.

$$\text{Jaccard Similarity}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

3. 유사도 함수

❖ Jaccard Similarity (자카드 유사도)

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?

- 사용자 A와 B가 공통으로 평가한 아이템: 7번방의 선물, 검은사제들, 극한직업, 엑시트, 파묘 (총 5개)
- 사용자 A와 B가 평가한 모든 아이템: 7번방의 선물, 검은사제들, 극한직업, 엑시트, 파묘, 국제시장 (총 6개)
- 자카드 유사도:

$$\text{Jaccard Similarity}(A, B) = \frac{5}{6} = 0.833$$

3. 유사도 함수

❖ 장단점

방법	장점	단점
MSD	- 간단하고 직관적 - 평가의 차이를 정확히 반영	- 사용자간의 평가 스케일이 다른 경우 정확한 반영이 안됨(주로 높은 점수를 주는 사용자와 주로 낮은 점수를 주는 사용자) => 패턴이 유사하더라도 점수 차이가 큼으로써 유사도가 낮게 나옴.
코사인 유사도	- 비교적 간단한 방법 - 스케일 불변성 (벡터의 크기와 무관)	- 사용자간의 평가 스케일 고려가 안됨. (패턴만 반영됨) => 패턴만 유사하면 유사도는 높게 나옴. - 결측값을 0으로 대체시 정보 손실
피어슨 상관계수	- 평가 스케일이 다른 사용자간의 유사도를 정확하게 측정 가능	- 아이템이 적은 경우 신뢰도가 떨어짐
자카드 유사도	- 이진 평가 데이터에 적합함	- 평점 데이터 활용 불가 - 아이템이 적은 경우 정확한 유사도 제공이 어려움

4. 사용자 기반 협업 필터링과 아이템 기반 협업 필터링 비교

사용자 기반 협업필터링	아이템 기반 협업 필터링
다른 사용자의 평가를 바탕으로 추천이 이루어지므로 정확도가 떨어질 수 있음	사용자의 자신의 평가 데이터를 활용하여 추천을 수행하기 때문에 더 관련성 높은 추천을 제공
shilling 공격에도 더 강함. * Shilling 공격 : 생산자가 의도적으로 본인 제품 선호도를 높게 작성하고 경쟁사 제품을 낮게 작성하는 경우	다양한 항목을 추천할 가능성이 높아 추천 목록의 다양성이 더 커질 수 있음.
추천의 이유를 명확하게 설명할 수 있습니다. ex) "이 영화를 봤기 때문에"라는 이유로 추천을 제공하는데, 이는 아이템 기반 방식으로 쉽게 설명됨.	이웃 사용자의 정보가 익명으로 처리되기 때문에 설명하기 어려움.

5. 이웃기반 방법의 장단점

❖ 장점

- 단순함과 직관적인 접근 방식
  - 구현과 디버깅이 쉽고, 특정 아이템이 추천된 이유를 설명하기도 용이함.
  - 특히, 아이템 기반 방법의 해석 가능성이 두드러짐.

❖ 단점

- 대규모 환경에서 오프라인 단계가 비현실적일 수 있음.
- 사용자 기반 방법의 오프라인 단계는 최소한 O(m^2)의 시간과 공간을 필요로 함.
  - 이는 수천만 명의 사용자 규모에서는 데스크탑 하드웨어로 처리하기에 너무 느리거나 공간이 많이 필요할 수 있음.
- 데이터가 희소할 경우, 이웃 기반 방법의 적용 범위가 제한될 수 있으며 유사도 계산이 어려워지는 문제가 발생함.

6. 모델 기반 협업 필터링

❖ 모델 기반의 협업 필터링이란 ?

- 모델 기반 협업 필터링은 데이터 마이닝과 머신 러닝 기법을 활용하여, 사용자와 아이템 간의 평가 패턴을 학습합니다.
- 이 방식은 추천 시스템에 필요한 일반화된 모델을 만들어 내고, 이 모델을 이용해 새로운 사용자나 아이템에 대해 빠르고 효율적으로 예측할 수 있게 함.

6. 모델 기반 협업 필터링

❖ 모델 기반의 협업 필터링 장단점

- 장점
  - 예측 성능이 뛰어남: 모델 기반 기법은 데이터를 학습하여 예측하므로 메모리 기반 방식보다 더 정교하고 정확한 추천을 생성
  - 일반화 가능: 데이터의 패턴을 학습한 후, 새로운 사용자나 아이템에 대해서도 예측이 가능하므로 확장성이 뛰어남.
  - 대규모 데이터 처리에 적합: 모델이 한 번 학습되면 예측 과정이 빠르기 때문에 대규모 데이터셋에서도 실시간 추천이 가능함.
- 단점
  - 학습 비용: 모델 학습 과정에서 상당한 계산 비용이 들며, 특히 딥러닝이나 행렬 분해 같은 고급 기법의 경우, 학습에 많은 시간이 소요될 수 있음
  - 데이터 의존성: 학습 데이터를 기반으로 모델이 만들어지므로, 학습 데이터의 품질이 모델 성능에 큰 영향을 미칩니다. 잘못된 학습 데이터가 들어가면 추천 품질이 떨어질 수 있음.
  - 복잡성 증가: 모델 기반 기법은 메모리 기반 방식에 비해 더 복잡하며, 모델 설정 및 튜닝에 시간이 많이 걸릴 수 있음.

6. 모델 기반 협업 필터링

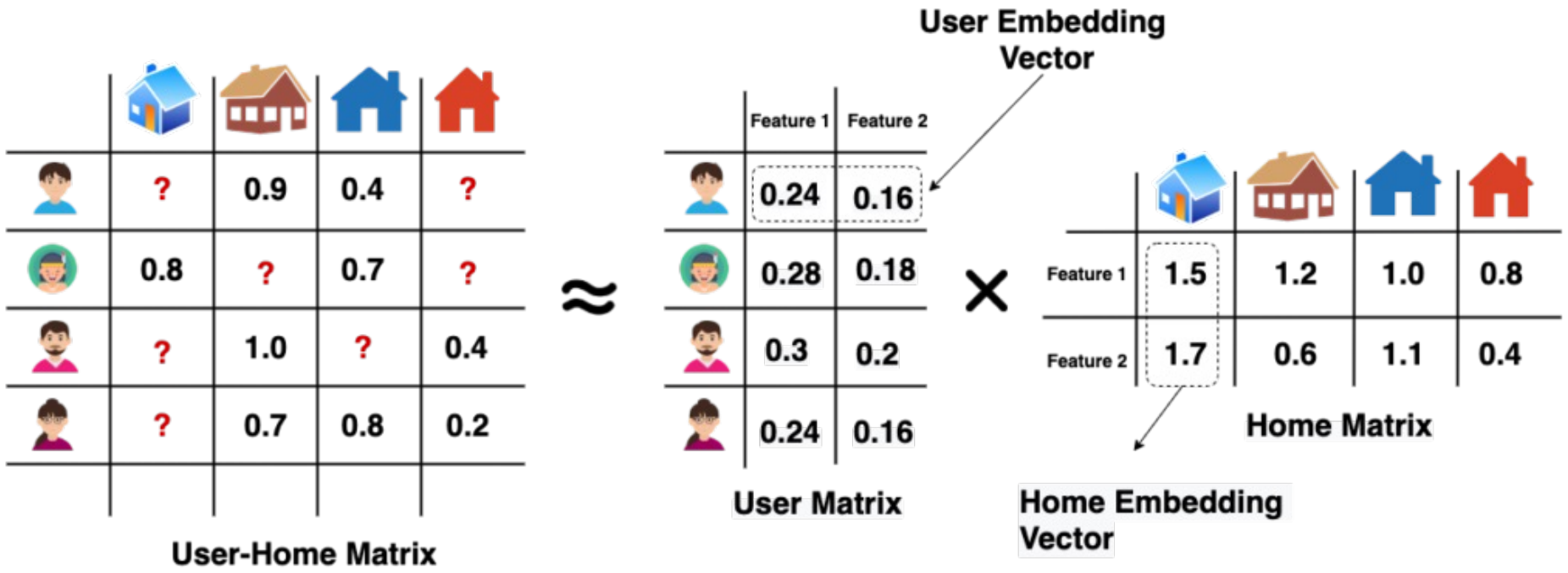
❖ 모델 기반의 협업 필터링은 언제 사용할까 ?

- 대규모 데이터셋을 다루는 경우
  - 모델이 한 번 학습된 후에는 매우 빠르게 예측을 수행할 수 있음.
- 데이터 희소성이 높은 경우
  - 행렬 분해와 같은 기법은 데이터 희소성을 잘 처리함.
- 복잡한 추천이 필요한 경우
  - 딥러닝 기반 기법은 사용자와 아이템 간의 복잡한 상호작용을 학습할 수 있으므로, 비선형적인 패턴을 포함한 복잡한 추천이 필요한 경우 적합함.

6. 모델 기반 협업 필터링

❖ Matrix Factorization (행렬분해)

- 사용자-아이템 평점 행렬을 저차원 행렬로 분해하여 추천을 생성하는 방법
- 평점 행렬의 희소성을 해결하고 잠재 요인(latent feature)을 추출하는 데 유용 (ex. feature1, feature2)



https://medium.com/@rebirth4vali/implementing-matrix-factorization-technique-for-recommender-systems-from-scratch-7828c9166d3c

6. 모델 기반 협업 필터링

❖ Matrix Factorization (행렬분해)

- 고객의 feature와 집의 feature가 비슷하다면 내적은 1이 될 것이고, 비슷하지 않다면 0이 됨.
- k 값이 클수록 원본 행렬을 잘 복원하지만 계산량은 늘어나고, k의 값이 작을수록 원본 행렬과의 오차는 커지지만 계산량은 줄어듦

	가격	접근성	인프라
고객1	?	?	?
고객2	?	?	?
고객3	?	?	?

	집1	집2	집3
가격	?	?	?
접근성	?	?	?
인프라	?	?	?

	집1	집2	집3
고객1	0.7	0.9	0.2
고객2	0.1	0.3	0.8
고객3	0.2	0.4	0.9

6. 모델 기반 협업 필터링

❖ Classification/Regression(분류/회귀) 방식

- Classification/Regression 방식은 콘텐츠 기반 추천 방식과 쉽게 융합이 가능
- 분류 방식 : 이진 또는 다중 클래스 레이블을 예측하는 방식
  - Ex) 사용자가 특정 영화를 좋아할지("좋아요" 또는 "싫어요")를 예측하는 문제는 이진 분류 문제로 해결 가능
- 알고리즘 예시: 결정 트리(Decision Trees), 랜덤 포레스트(Random Forest), 서포트 벡터 머신(SVM), 로지스틱 회귀(Logistic Regression) 등

idx	X1(가격)	X2(인프라)	X3(접근성)	Y(likes)
1	2000	3	5	1
2	3000	4	3	0
3	5000	5	3	?

$$Y = a * x1 + b + x2 + c * x3 + d$$

6. 모델 기반 협업 필터링

❖ Classification/Regression(분류/회귀) 방식

- Classification/Regression 방식은 콘텐츠 기반 추천 방식과 쉽게 융합이 가능
- 회귀 방식 : 연속적인 수치를 예측
  - Ex) 사용자가 특정 아이템에 대해 줄 평점(예: 1~5점)을 예측하는 문제는 회귀로 해결 가능
- 알고리즘 예시: 선형 회귀(Linear Regression), 다항 회귀(Polynomial Regression), 결정 트리 회귀(Decision Tree Regression)

idx	X1(가격)	X2(인프라)	X3(접근성)	Y(rating)
1	2000	3	5	5
2	3000	4	3	4
3	5000	5	3	?

$$Y = a * x1 + b + x2 + c * x3 + d$$

6. 모델 기반 협업 필터링

❖ Classification/Regression 방식의 장점

1. 콘텐츠 기반 추천의 확장성
   - Classification/Regression 방식은 사용자와 아이템 간의 관계를 학습하는 데 강력한 도구로 작동하므로, 콘텐츠 기반 추천에 더 정교한 예측 모델을 추가할 수 있음.
   - 특히, 아이템의 속성 정보가 풍부한 경우 이 방식은 매우 유용함.
2. 유연성
   - 다양한 형태의 데이터에 쉽게 적용될 수 있음.
   - 텍스트 데이터, 카테고리 데이터, 연속 데이터 등 다양한 특징을 다룰 수 있으며, 사용자 선호도나 평점 예측에 적합함.
3. 콜드 스타트 문제 완화
   - 분류 및 회귀 방식은 사용자가 직접 평점을 남기지 않더라도, 아이템의 특성 정보만으로도 예측을 수행할 수 있으므로 콜드 스타트 문제를 완화하는 데 도움이 됨.

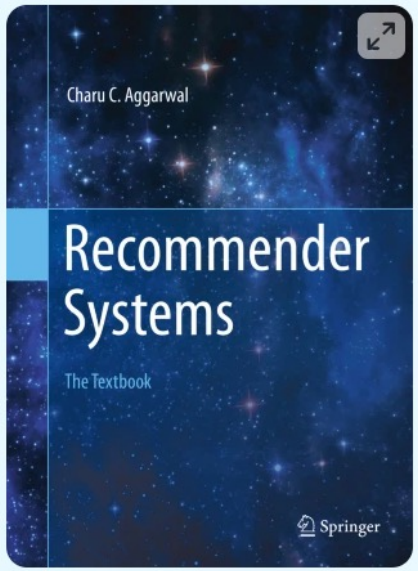
Mini Quiz

❖ 아이템 기반의 협업 필터링으로 "사용자4의 검은사제들"에 대한 예측 평점을 구해보세요.

- 검은사제들과 유사한 아이템으로 정의할 유사도 점수 기준은 0.9입니다.
- Hint : slide 11을 "열(column)"기준으로 구해보세요.

	7번방의 선물	검은사제들	극한직업	엑시트	파묘	국제시장
사용자1	5	1	3	2	2	5
사용자2	5	2	2	2	2	?
사용자3	1	5	4	4	5	1
사용자4	2	?	4	5	4	1
유사도 점수	0.56		0.92	0.92	0.97	0.49

참고자료



Recommender Systems

The Textbook

Textbook | © 2016

Access provided by Korea University

Download book PDF

Download book EPUB

https://link.springer.com/book/10.1007/978-3-319-29659-3

이 문서의 외부 유출 및 공유를 금합니다.

DATA CREATOR CAMP

2024 데이터 크리에이터 캠프

감사합니다.

과학기술정보통신부 NIA 지능정보원 한국지능정보사회진흥원