boostcamp

Collaborative Filtering

고 지 형

Collaborative Filtering, CF

유저, 상품 간 상호작용을 활용한 추천 시스템

- 유저가 상품에 매긴 평점(rating) 기반의 추천
- 이웃한(유사한) 유저 또는 상품을 활용한다는 점에서 Neighbor-based CF 라고도 부름
- 유저와 상품에 대한 어떠한 속성 정보(attributes)도 필요하지 않음

User-based CF

- *'두 유저가 비슷한 취향을 가졌다면 앞으로도 비슷한 취향을 가질 것이다.'*
- 여러 유저들이 한 상품에 매긴 평점을 바탕으로 추천
- 유저 간 유사도를 바탕으로 상품 추천

Item-based CF

- *'유저의 취향은 앞으로도 유지될 것이다'*
- 상품 간 유사도를 바탕으로 상품 추천
- 한 유저가 여러 상품에 매긴 평점을 바탕으로 추천
- 유저 자신의 평점을 활용한다는 점에서 안정적인 추천
- User-based CF보다 더 많이 사용

Collaborative Filtering

boostcamp

추천 과정

1. 유저(상품) 간 유사도 측정
2. 유사한 k명(개)의 유저(상품)의 피어 그룹 선별
3. 피어 그룹을 활용해 특정 상품에 대한 평점을 추정
4. 추정한 평점을 바탕으로 그럴듯한 상품을 추천

전제

- m명의 유저, n개의 상품으로 이루어진 희소 행렬(sparse matrix)

Item-Id ⇒ / User-Id ⇓	1	2	3	4	5	6
1	7	6	7	4	5	4
2	6	7	?	4	3	4
3	?	3	3	1	1	?
4	1	2	2	3	3	4
5	1	?	1	2	3	3

m

n

User-based Collaborative Filtering

boostcamp

유저 간 유사도 측정

- 기준 유저 u와 그 외 유저 간 유사도를 측정
- 유저 u와 v가 공통적으로 평가한 상품($I_u \cap I_v$)에 대한 평점을 바탕으로 피어슨 상관계수 측정
- 피어슨 상관계수에 활용되는 표본 평균(μ_n): 유저가 지금까지 평가한 모든 상품에 대한 평점의 평균

$$\text{Sim}(u,v) = \text{Pearson}(u,v) = \frac{\sum_{k \in I_u \cap I_v}(r_{uk} - \mu_u) \cdot (r_{vk} - \mu_v)}{\sqrt{\sum_{k \in I_u \cap I_v}(r_{uk} - \mu_u)^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v}(r_{vk} - \mu_v)^2}}$$

$$\mu_u = \frac{\sum_{k \in I_u} r_{uk}}{|I_u|} \quad \forall u \in \{1 \dots m\}$$

User-based Collaborative Filtering

boostcamp

평점 추정: 유저 u의 상품 j에 대한 평점 예측

(1) 피어 그룹 선별

- 측정한 유사도를 바탕으로 유저 u와 유사한 K명의 유저를 피어 그룹으로 선발
- 피어 그룹의 유저는 상품 j에 대해 평가한 이력이 있어야 함
- 피어 그룹: $P_u(j)$

(2) 정규화

- 어떤 유저는 대부분의 상품에 높은 점수를, 어떤 유저는 낮은 점수를 부여할 수 있음 => 편중 현상
- 편중된 평점 추정을 방지하기 위해, 평점 추정에 앞서 각 유저들의 상품 j에 대한 평점을 스케일링(S_{uj})
- 각 유저들의 상품 j에 대한 평점에 지난 모든 상품에 대한 평점 평균을 차감

(3) 평점 추정

- 피어그룹 유저들의 상품 j에 대한 평점을 가중 평균하여 유저 u의 평점 추정
- 가중 평균에 활용되는 가중치는 유사도
- 유저 u의 평점 평균을 더하여 역정규화

$$s_{uj} = r_{uj} - \mu_u \quad \forall u \in \{1 \dots m\}$$

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

User-based Collaborative Filtering

boostcamp

예시: 유저3에 대한 상품 추천

Table 2.1: User-user similarity computation between user 3 and other users

Item-Id ⇒ / User-Id ⇓	1	2	3	4	5	6	Mean Rating	Cosine$(i,3)$ (user-user)	Pearson$(i,3)$ (user-user)
1	7	6	7	4	5	4	5.5	0.956	0.894
2	6	7	?	4	3	4	4.8	0.981	0.939
3	?	3	3	1	1	?	2	1.0	1.0
4	1	2	2	3	3	4	2.5	0.789	-1.0
5	1	?	1	2	3	3	2	0.645	-0.817

추정할 평점

1. 유사도 측정

유사도가 높은 유저1, 유저2 => 피어 그룹으로 선발

2. 평점 추정

정규화를 거친 유저1과 유저2의 평점을 가중 평균한 뒤,
유저3의 평균 평점을 더해 평점을 추정

$$\hat{r}_{31} = 2 + \frac{1.5 * 0.894 + 1.2 * 0.939}{0.894 + 0.939} \approx 3.35$$

$$\hat{r}_{36} = 2 + \frac{-1.5 * 0.894 - 0.8 * 0.939}{0.894 + 0.939} \approx 0.86$$

Item-based Collaborative Filtering

boostcamp

상품 간 유사도 측정

- 상품에 여러 유저들이 매긴 평점을 바탕으로 유사도를 측정
- 기준 상품 i와 그 외 상품 j 모두 평가한 적 있는 유저($U_i \cap U_j$)의 평점 정보 활용

(1) 정규화

- 유사도를 측정하기에 앞서서 유저별 평점에 대한 정규화를 진행
- User-based CF에 활용된 정규화와 동일(S_{uj})

(2) 유사도 측정

- 기준 상품 i와 그 외 상품 간 유사도를 측정
- 수정된 코사인 유사도(Adjusted Cosine) 활용
  - 유사도 측정 이전에 정규화를 취했기 때문에 '수정된'을 덧붙임
  - Item-based CF에서 피어슨 상관계수보다 일반적으로 성능이 좋음

$$\text{AdjustedCosine}(i,j) = \frac{\sum_{u \in U_i \cap U_j} s_{ui} \cdot s_{uj}}{\sqrt{\sum_{u \in U_i \cap U_j} s_{ui}^2} \cdot \sqrt{\sum_{u \in U_i \cap U_j} s_{uj}^2}}$$

$$s_{uj} = r_{uj} - \mu_u \quad \forall u \in \{1 \dots m\}$$

Item-based Collaborative Filtering

boostcamp

평점 추정: 유저 u의 상품 t에 대한 평점 예측

(1) 피어 그룹 선별

- 측정한 유사도를 바탕으로 상품 t와 유사한 K개의 상품을 피어 그룹으로 선발
- 피어 그룹의 상품은 유저 u가 평점 매긴 적 있는 상품이어야 함
- 피어 그룹: $Q_t(u)$

(2) 평점 추정

- 피어 그룹 각각의 상품 j에 대한 평점을 가중 평균하여 유저 u의 상품 t에 대한 평점을 추정
- 가중 평균에 활용되는 가중치는 수정된 코사인 유사도

$$\hat{r}_{ut} = \frac{\sum_{j \in Q_t(u)} \text{AdjustedCosine}(j,t) \cdot r_{uj}}{\sum_{j \in Q_t(u)} |\text{AdjustedCosine}(j,t)|}$$

Item-based Collaborative Filtering

boostcamp

예시: 유저3에 대한 상품 추천

추정할 평점

Item-Id ⇒ / User-Id ⇓	1	2	3	4	5	6
1	1.5	0.5	1.5	-1.5	-0.5	-1.5
2	1.2	2.2	?	-0.8	-1.8	-0.8
3	?	1	1	-1	-1	?
4	-1.5	-0.5	-0.5	0.5	0.5	1.5
5	-1	?	-1	0	1	1
Cosine$(1, j)$ (item-item)	1	0.735	0.912	-0.848	-0.813	-0.990
Cosine$(6, j)$ (item-item)	-0.990	-0.622	-0.912	0.829	0.730	1

1. 정규화

수정된 코사인 유사도 측정에 앞서 사전에 평점 정규화

2. 유사도 측정

상품 1에 대한 피어 그룹: 상품 2, 상품 3

상품 6에 대한 피어 그룹: 상품 4, 상품 5

$$\hat{r}_{31} = \frac{3 * 0.735 + 3 * 0.912}{0.735 + 0.912} = 3$$

$$\hat{r}_{36} = \frac{1 * 0.829 + 1 * 0.730}{0.829 + 0.730} = 1$$

3. 평점 추정

피어 그룹 내 상품 평점과 각 유사도를 활용한 가중 평균

User-based vs. Item-based

boostcamp

무엇이 더 좋나?

- 일반적으로 Item-based CF가 더 좋다고 알려짐: 유저 자신의 평점 기록을 활용한 추천을 하기 때문
- 하지만, 근본적으로 핵심은 데이터의 질
- 상품의 다양성, 유저의 풍부한 평가 기록 등이 밑바탕 되어야 두 방법이 모두 잘 작동
- 데이터가 부족하면, 창의성, 다양성, 우연성이 결여된 '지루한' 추천이 이루어짐

설명력

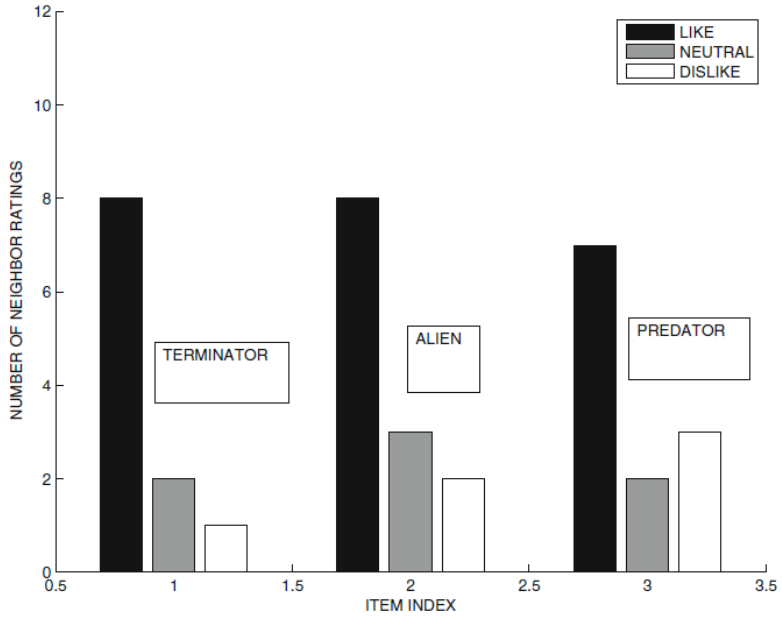
Item-based CF

- 유저가 과거에 평점을 매겼던 상품을 추천의 근거로 제시할 수 있어 설명력이 뛰어남

$$Because\ you\ watched\ \text{“}Secrets\ of\ the\ Wings,\text{”}\ [the\ recommendations\ are]\ \langle List\rangle\ .$$

User-based CF

- 유저와 유사한 유저 그룹의 분포를 보여줌으로써 추천의 근거를 제시
- 유저 자신의 취향은 반영하지 않아 Item-based CF에 비해 부족한 설명



감사합니다