CHAPTER 2. NEIGHBORHOOD-BASED COLLABORATIVE FILTERING

Chapter 2.2 Key Properties of Rating Matrices, p.32-33

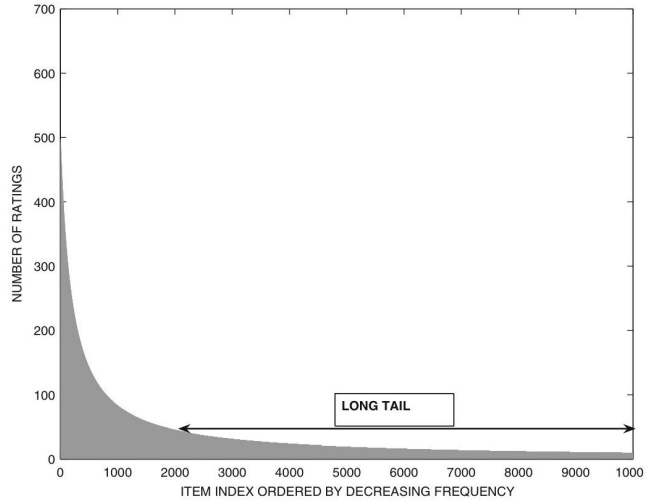

Figure 2.1: The long tail of rating frequencies

x-axis: 판매/추천 빈도, y-axis: popularity(얼마나 빈번하게 rating 되었는지)

high-frequency items: little profit for the merchant.
lower-frequency items: larger profit margins.

판매량과 이익이 반비례하는 이유

- 판매자-공급자 협상 지위의 관점: 판매자 - 온라인 플랫폼(ex. amazon, coupang) - 의 경우, 판매량이 많은 제품(인기 제품)은 공급자의 자체 경쟁력이 높기때문에, **공급자가 판매자 대비 협상에서 우위**를 점할 수 있음. 따라서 판매자에 돌아가는 수익이 적을 수 있다.(margin ↓)
  반면, 추천 빈도수가 낮거나 판매량이 적은 비인기 제품은 대체로 공급자 자체 경쟁력이 비교적 낮기 때문에, **판매자가 협상에서 우위**를 점할 수 있음. 따라서 온라인 플랫폼 기업에 돌아가는 마진이 더 크다.
- 재고 관리의 관점: 재고를 남기는 것 자체가 비용. 재고 비용을 줄이는 관점에서 보면, 판매량이 낮은 롱 테일 제품을 판매하여 악성 재고를 줄일 수 있음. 롱 테일 제품을 처리하는 것이 회사의 주요 이슈 중 하나

Chapter 2.3.1 User-Based Neighborhood Models, p.37

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

mean-centering process가 더 좋은 예측을 해주지만, allowed ratings에서 벗어난다는 단점이 있다. 본문의 예시에서는 $\widehat{r}_{3,6} = 0.85$ 값을 outside the range of allowed ratings라고 본다. 그렇다면 allowed ratings를 벗어난다는 것이 어떠한 range를 기준으로 말하는 것인가?

→ rating을 1~7점까지 매기는데, 0.85는 이 범위에서 벗어난다는 의미인 듯.

Chapter 2.3.4 Comparing User-Based and Item-Based Methods, p.43-44

user-based	**item-based**
serendipity ↑ : recommend some novel items	accuracy ↑ : target user's own ratings are used
provide a **concrete reason X** : peer group is a set of anonymous users	provide a **concrete reason** : because you watched.
	more stable : #(users) > #(items) : new users added more frequently

item-based method는 상대적으로 serendipity ↓, 추천 item의 diversity ↓
: 기존에 이미 평가를 했던 items 기반으로 추천을 하기 때문에,
기존에 평가했던 item과 같은 카테고리의 items가 추천이 될 확률이 크다.
ex. 텐트를 구매했다면, 캠핑 용품 카테고리의 캠핑 의자 등의 item을 추천

반면, user-based method는 user 개인의 패턴과 잠재적인 특성이 반영되어 추천을 한다.
ex. user1(target user의 neighbor로 가정)이 텐트와 기저귀를 구매했다면, 이미 텐트를 구매했던 target user에게 기저귀가 추천될 수 있다. accuracy는 떨어질 수 있지만, 추천의 다양성은 제공된다.

2.3.5 Strengths and Weaknesses of Neighborhood-Based Methods, p.44

강점

- 직관적이고 설명하기 쉬움. Item-Based는 관련 아이템 기반, User-Based는 관련 유저 기반으로 추천된 결과. 반면 머신러닝 기반 Model-Based(3장 내용)는 그렇지 않음
- 새로운 유저가 아이템이 추가되어도 stable

약점

- **Offline phase에서 시간 복잡도가 커짐**. 유저가 m개라면 m^2의 조합을 계산해야 하기 때문
    - Offline phase: User 간 / Item 간 유사도를 계산하는 과정
    - Online phase: 유사도에 기반하여, top-k set을 이용하여 가중 평균으로 점수를 예측하는 과정
- Sparsity Problem: User-Based에서 나의 k-neighbor 중 타겟 아이템을 평가한 사람이 한 명도 없다면, rating을 예측할 수 없는 문제 발생. 다만 한편으로는 이 현상이 좋지 않은 추천이라는 의미가 될 수 있음