

빅데이터 분석 프로그래밍

Application

Week-05. Recommender System Basic

Jungwon Seo, 2022-Fall

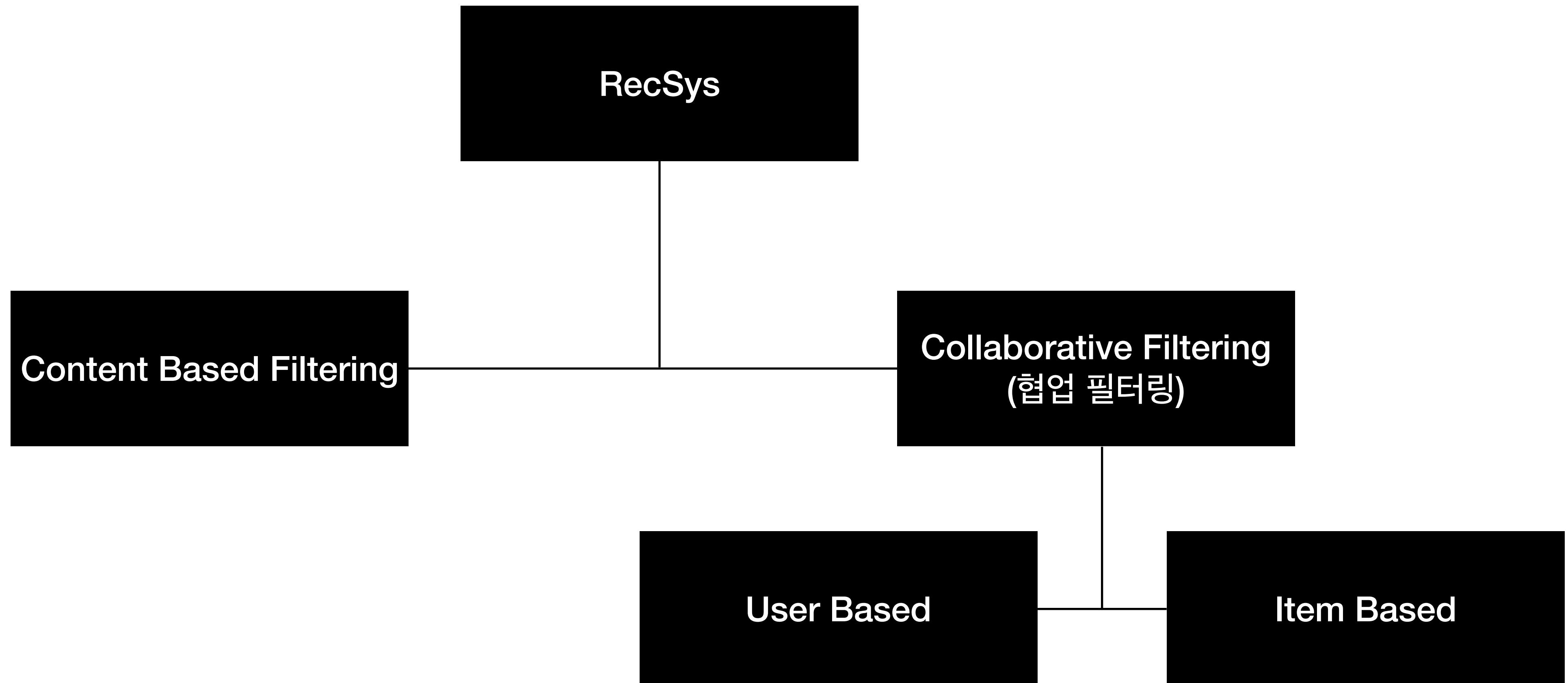
추천시스템

내가 이걸 좋아한다면?

- 추천 시스템은 정보 필터링 기술의 일종으로,
특정 사용자가 관심을 가질만한 정보를 추천하는 것
 - 예1: 온라인 쇼핑몰 상품 추천, 넷플릭스 콘텐츠 추천, 구글 광고 추천
 - 예2: 금융 상품 추천, (펀드, 적금 등)
- How?
 - User Based Collaborative Filtering
 - Item Based Collaborative Filtering
 - Content based filtering

추천시스템

추천 시스템의 종류



User Based Collaborative Filtering

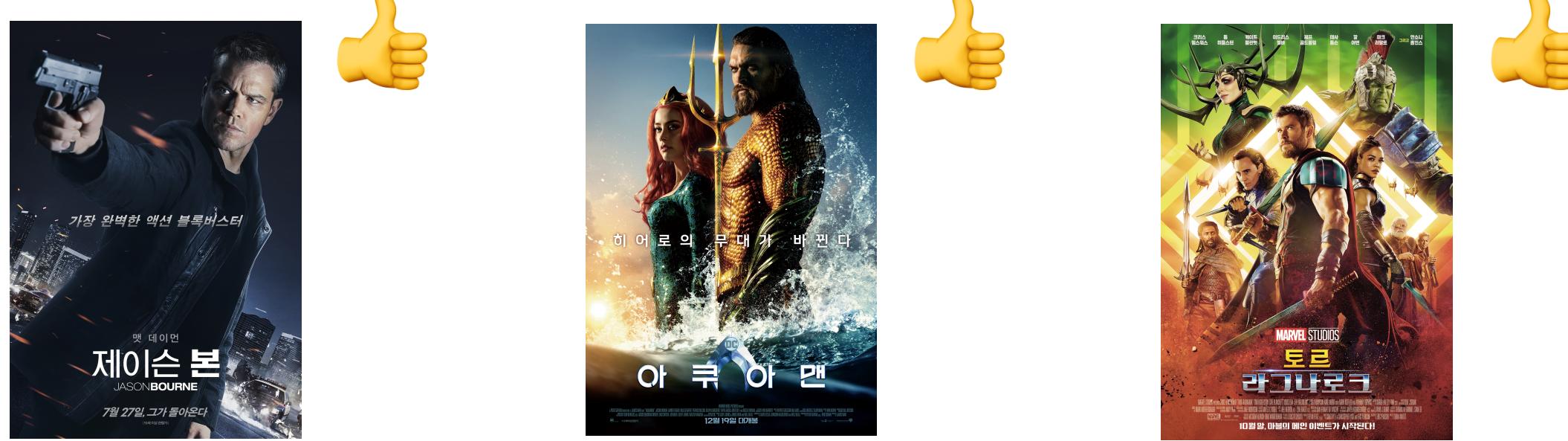
사용자의 행동을 기반으로 쌓은 데이터를 활용

- 나와 비슷한 다른 사용자를 찾기
 - item과의 interaction들을 벡터 형태로 비교
- 데이터 요구 사항
 - 사용자의 Interaction 기록이 있어야함
- 데이터 수집 방식
 - 소셜 미디어의 “좋아요” 기능
 - 쇼핑몰의 장바구니 기능
 - 구글 광고 클릭 기능
- 이러한 데이터를 기반으로 사용자들을 Grouping 가능

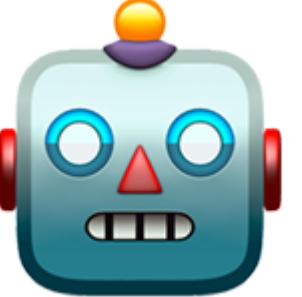
User Based Collaborative Filtering

Example: 영화 추천

User1



User2



User3



User Based Collaborative Filtering

Example: User1과 User3는 비슷한 취향이다!



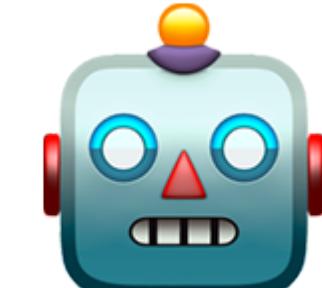
User Based Collaborative Filtering

비슷하다는 것을 어떻게 계산할 수 있을까?

	제이슨본	아쿠아맨	토르	스파이더맨	건축학개론	암살
	1	1	1	0	0	0
	0	0	0	1	1	1
	1	1	0	0	0	0

User Based Collaborative Filtering

한 유저들을 각각의 Vector로 보고 영화는 축으로 => Cosine Similarity

	제이슨본	아쿠아맨	도르	스파이더맨	건축학개론	암살	유사도
	1	1	1	0	0	0	1
	0	0	0	1	1	1	0
	1	1	0	0	0	0	0.81

Cosine Similarity

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

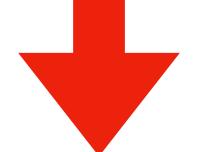
Item Based Collaborative Filtering

비슷한 영화를 추천해주는 건 어떨까?

- 영화가 비슷하다는 것은?
- Collaborative Filtering 기반에서는, 영화를 Grouping
- 예를들어
 - User1은 어벤져스를 보고, 아바타를 본상태
 - User2는 어벤져스만 본 상태
 - 어벤져스와 아바타는 비슷한 부류의 영화군!

Item Based Collaborative Filtering

비슷하다는 것을 어떻게 계산할 수 있을까? Transpose



	스파이더맨	아쿠아맨	토르	제이슨본	존윅	아저씨
😊	1	1	1	0	0	0
🤖	0	0	0	1	1	1
😈	1	1	0	0	0	0
유사도	0.7	0.7	1	0	0	0

Collaborative Filtering

언제 User Based를 사용하고, 언제 Item Based를 사용해야하나?

- 성능이 좋은 방식으로
- 일반적으로 데이터의 interaction 비율에 따라서 성능의 차이가 나는 경우가 존재
 - Item당 interaction 수가 많은 경우 item-based
 - User당 interaction 수가 많은 경우 user-based
- 영화: 사용자가 영화 평가를 많이 하면 100개 정도겠지만, 영화 하나는 수십만, 수백만의 평가를 갖고 있음
 - 사용자들을 표현할 수 있는 feature 수보다, 영화를 표현할 수 있는 feature수가 많다는 뜻

이전에 아무도 선택하지 않은 영화는
어떻게 추천 할 수 있을까?

Content Based Filtering

“진짜” 비슷한 영화를 추천해주는 건 어떨까?

- 영화가 비슷하다는 것은?
- Content Based Filtering 기반에서는, 영화의 Feature들을 기반으로 유사도를 계산
- 예를들어, 감독, 장르, 주연배우, 언어 등
 - 더나아가 자연어 처리를 활용한다면 줄거리
 - 마찬가지로 유사도는 코사인 유사도
- 검색엔진과 유사
 - 검색엔진: query와 document의 유사도
 - 추천시스템: item과 다른 item들과의 유사도

E.O.D