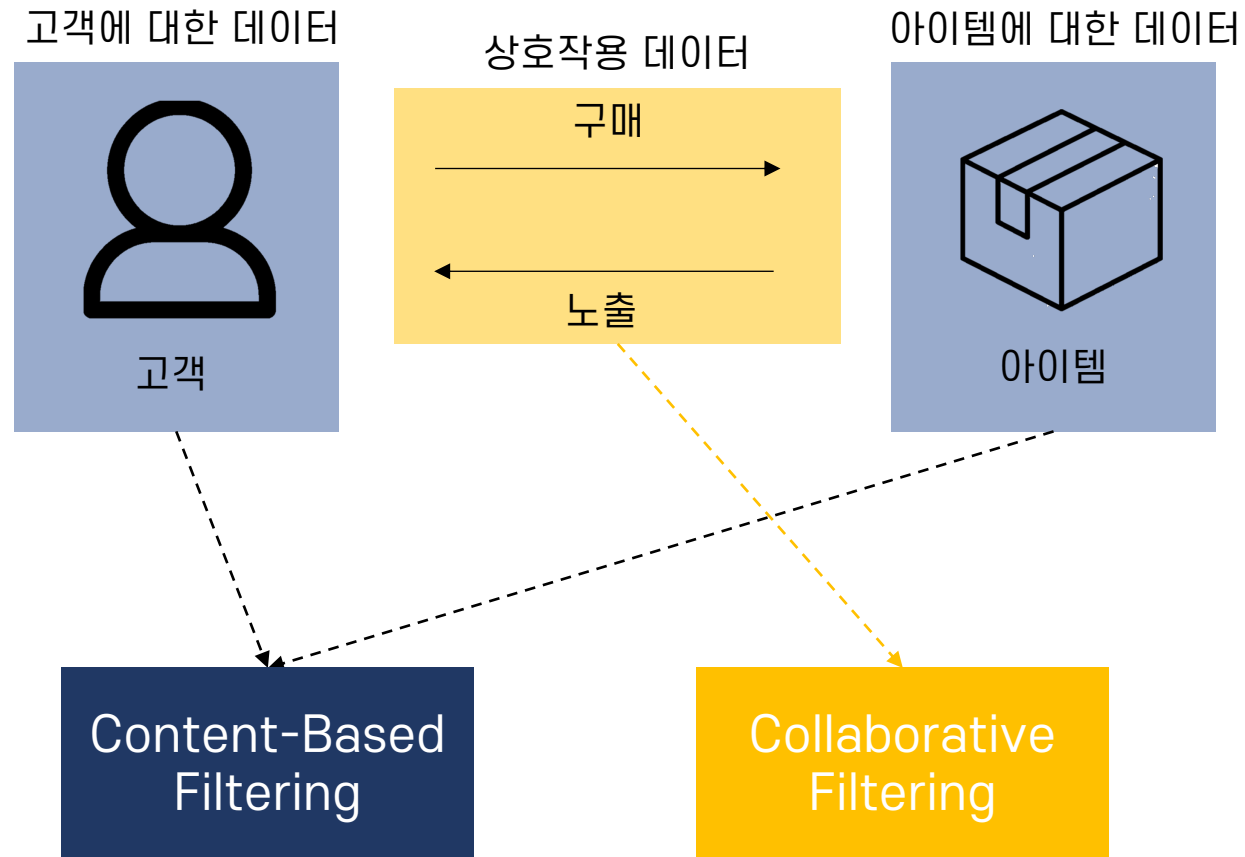


협업 필터링

<Collaborative Filtering>

추천 시스템의 알고리즘

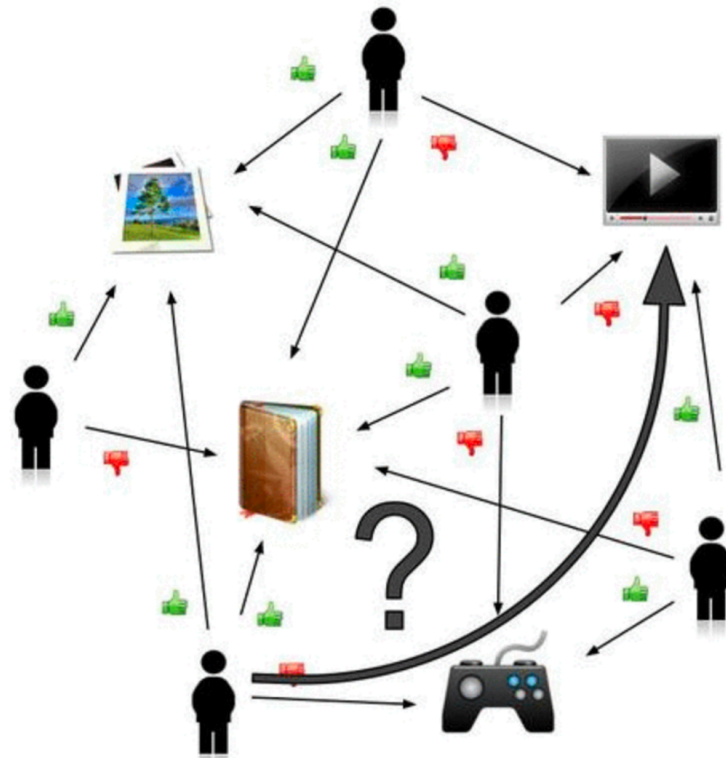
추천 시스템은 크게 콘텐츠 기반 추천과 협업 필터링 기반 추천으로 나뉘어짐



협업 필터링이란?

핵심 아이디어

만약 두 명의 사용자가 유사한 관심사를 가지고 있다면,
그들은 미래에도 유사한 취향을 가질 것이다.



CF 알고리즘의 데이터, 상호작용 정보

	user_id	movie_id	rating	rated_at
0	1	2	3.5	1112486027
175	2	3	4.0	974820889
236	3	1	4.0	944919407
423	4	6	3.0	840879227
424	4	10	4.0	840878922
451	5	2	3.0	851527569
517	6	1	5.0	858275452
518	6	3	3.0	858275558
519	6	7	5.0	858275558
541	7	3	3.0	1011208463
542	7	7	3.0	1011208220
817	8	1	4.0	833981871
818	8	3	5.0	833981733
819	8	6	3.0	833982631
820	8	10	4.0	833981834
922	10	1	4.0	943497887

고객이 아이템에게
어떠한 액션을 취했는지에 대한 정보

CF 알고리즘의 데이터, 상호작용 정보

	user_id	movie_id	rating	rated_at
0	1	2	3.5	1112486027
175	2	3	4.0	974820889
236	3	1	4.0	944919407
423	4	6	3.0	840879227
424	4	10	4.0	840878922
451	5	2	3.0	851527569
517	6	1	5.0	858275452
518	6	3	3.0	858275558
519	6	7	5.0	858275558
541	7	3	3.0	1011208463
542	7	7	3.0	1011208220
817	8	1	4.0	833981871
818	8	3	5.0	833981733
819	8	6	3.0	833982631
820	8	10	4.0	833981834
922	10	1	4.0	943497887

고객이 아이템에게
어떠한 액션을 취했는지에 대한 정보

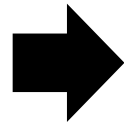
예시)

→ 6번 고객이 7번 영화에 평점 5점을 부여하였음

CF 알고리즘의 데이터, 상호작용 정보

	user_id	movie_id	rating	rated_at
0	1	2	3.5	1112486027
175	2	3	4.0	974820889
236	3	1	4.0	944919407
423	4	6	3.0	840879227
424	4	10	4.0	840878922
451	5	2	3.0	851527569
517	6	1	5.0	858275452
518	6	3	3.0	858275558
519	6	7	5.0	858275558
541	7	3	3.0	1011208463
542	7	7	3.0	1011208220
817	8	1	4.0	833981871
818	8	3	5.0	833981733
819	8	6	3.0	833982631
820	8	10	4.0	833981834
922	10	1	4.0	943497887

고객이 아이템에게
어떠한 액션을 취했는지에 대한 정보

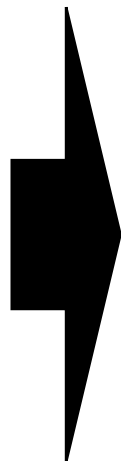


이러한 데이터의 포맷에서는
고객과 아이템 간 특징들을 파악하기 어려움

USER-ITEM Matrix

	user_id	movie_id	rating	rated_at
0	1	2	3.5	1112486027
175	2	3	4.0	974820889
236	3	1	4.0	944919407
423	4	6	3.0	840879227
424	4	10	4.0	840878922
451	5	2	3.0	851527569
517	6	1	5.0	858275452
518	6	3	3.0	858275558
519	6	7	5.0	858275558
541	7	3	3.0	1011208463
542	7	7	3.0	1011208220
817	8	1	4.0	833981871
818	8	3	5.0	833981733
819	8	6	3.0	833982631
820	8	10	4.0	833981834
922	10	1	4.0	943497887

PIVOT

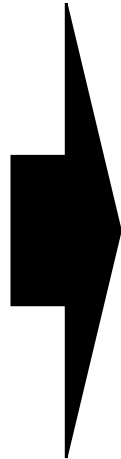


movie_id	1	2	3	6	7	10
user_id						
1	NaN	3.5	NaN	NaN	NaN	NaN
2	NaN	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	3.0	NaN	4.0
5	NaN	3.0	NaN	NaN	NaN	NaN
6	5.0	NaN	3.0	NaN	5.0	NaN
7	NaN	NaN	3.0	NaN	3.0	NaN
8	4.0	NaN	5.0	3.0	NaN	4.0
10	4.0	NaN	NaN	NaN	NaN	NaN

행(user)과 열(item)으로 정렬한 행렬

USER-ITEM Matrix

	user_id	movie_id	rating	rated_at
0	1	2	3.5	1112486027
175	2	3	4.0	974820889
236	3	1	4.0	944919407
423	4	6	3.0	840879227
424	4	10	4.0	840878922
451	5	2	3.0	851527569
517	6	1	5.0	858275452
518	6	3	3.0	858275558
519	6	7	5.0	858275558
541	7	3	3.0	1011208463
542	7	7	3.0	1011208220
817	8	1	4.0	833981871
818	8	3	5.0	833981733
819	8	6	3.0	833982631
820	8	10	4.0	833981834
922	10	1	4.0	943497887



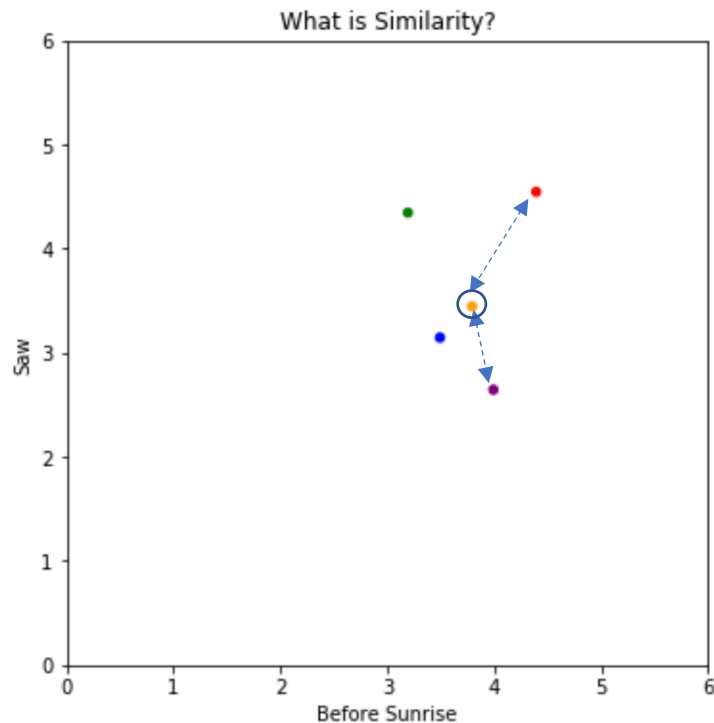
유사도						
movie_id	1	2	3	6	7	10
user_id						
1	NaN	3.5	NaN	NaN	NaN	NaN
2	NaN	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	3.0	NaN	4.0
5	NaN	3.0	NaN	NaN	NaN	NaN
6	5.0	NaN	3.0	NaN	5.0	NaN
7	NaN	NaN	3.0	NaN	3.0	NaN
8	4.0	NaN	5.0	3.0	NaN	4.0
10	4.0	NaN	NaN	NaN	NaN	NaN

유사도를 통해 우리는 어떤 영화끼리 비슷한지를
알 수 있음

협업 필터링 연산의 핵심 : 유사도

핵심 아이디어

만약 두 명의 사용자가 유사한 관심사를 가지고 있다면,
그들은 미래에도 유사한 취향을 가질 것이다.



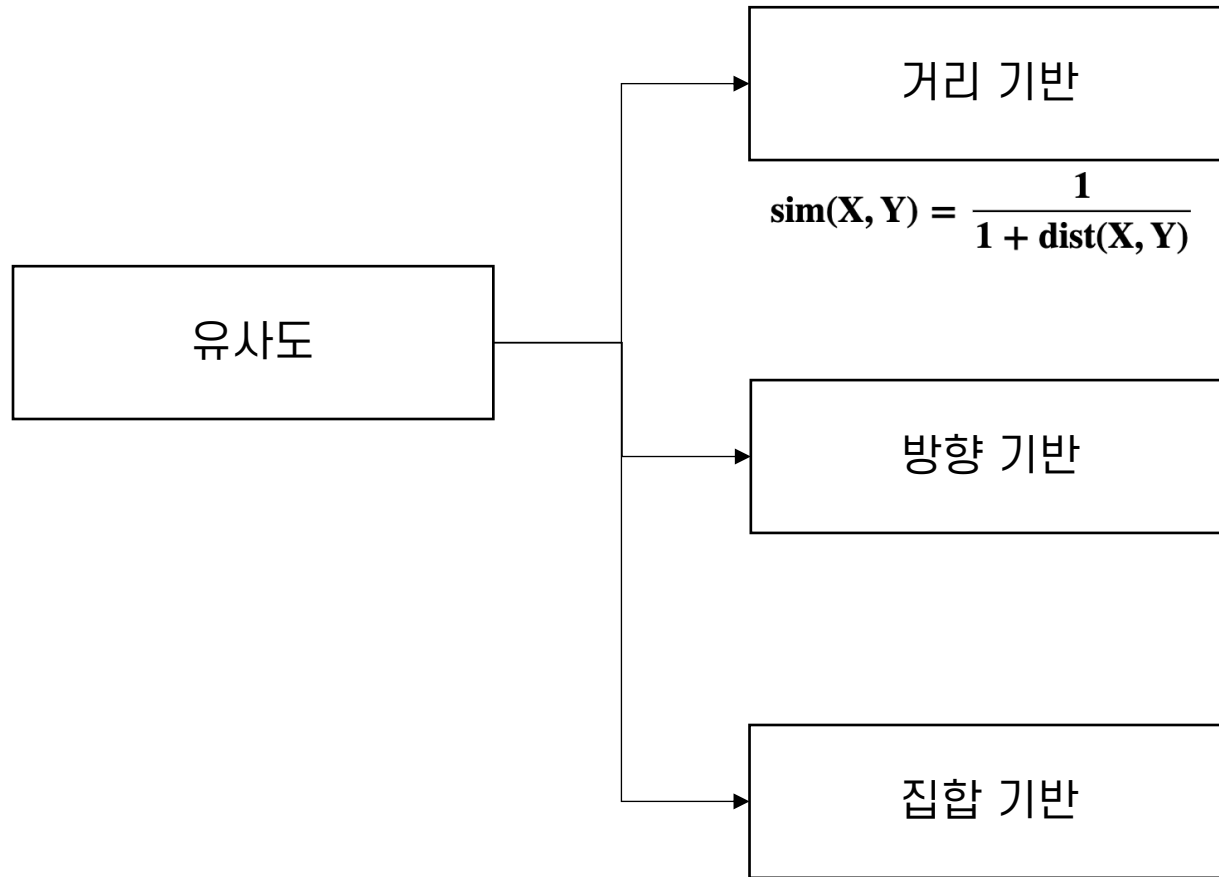
유사도 계산

노란색 유저는

* 빨간색 유저와 유사할까?

* 보라색 유저와 유사할까?

유사도의 다양한 기준들



맨해튼 유사도

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^n (|x_i - y_i|)$$

유클리디안 유사도

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

코사인 유사도

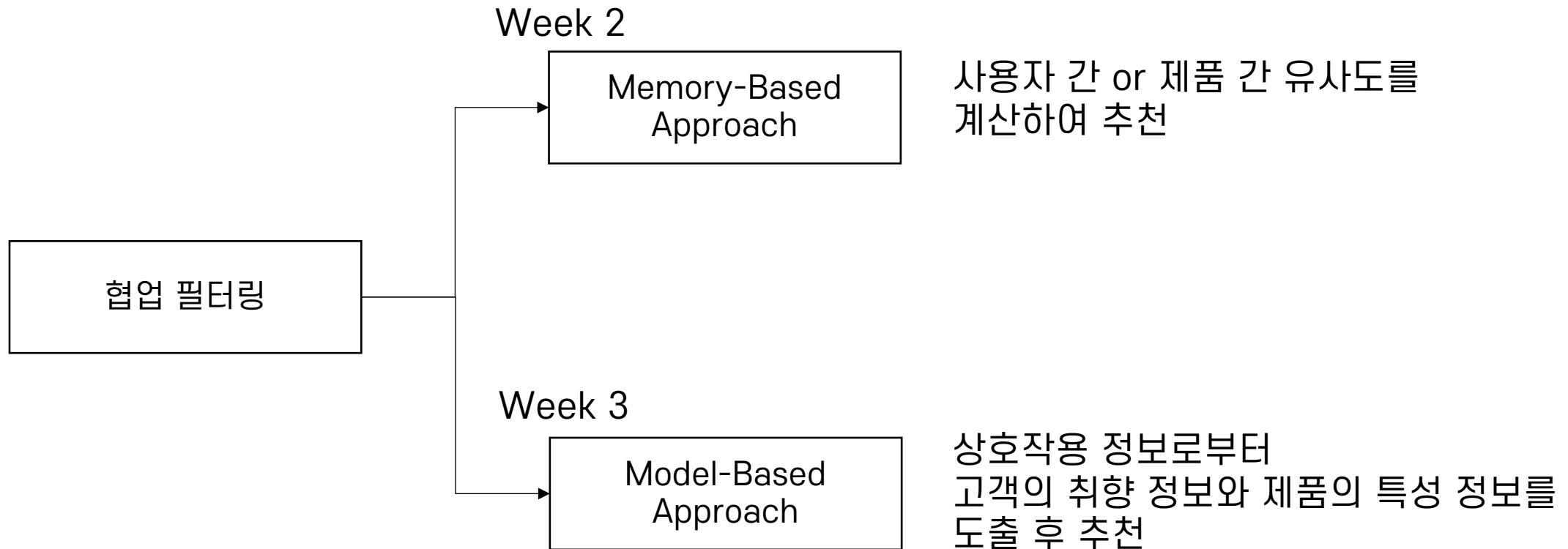
$$\text{sim}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

자카드 유사도

$$\text{sim}(\mathbf{X}, \mathbf{Y}) = \frac{(\mathbf{X} \cap \mathbf{Y})}{(\mathbf{X} \cup \mathbf{Y})}$$

협업 필터링의 종류

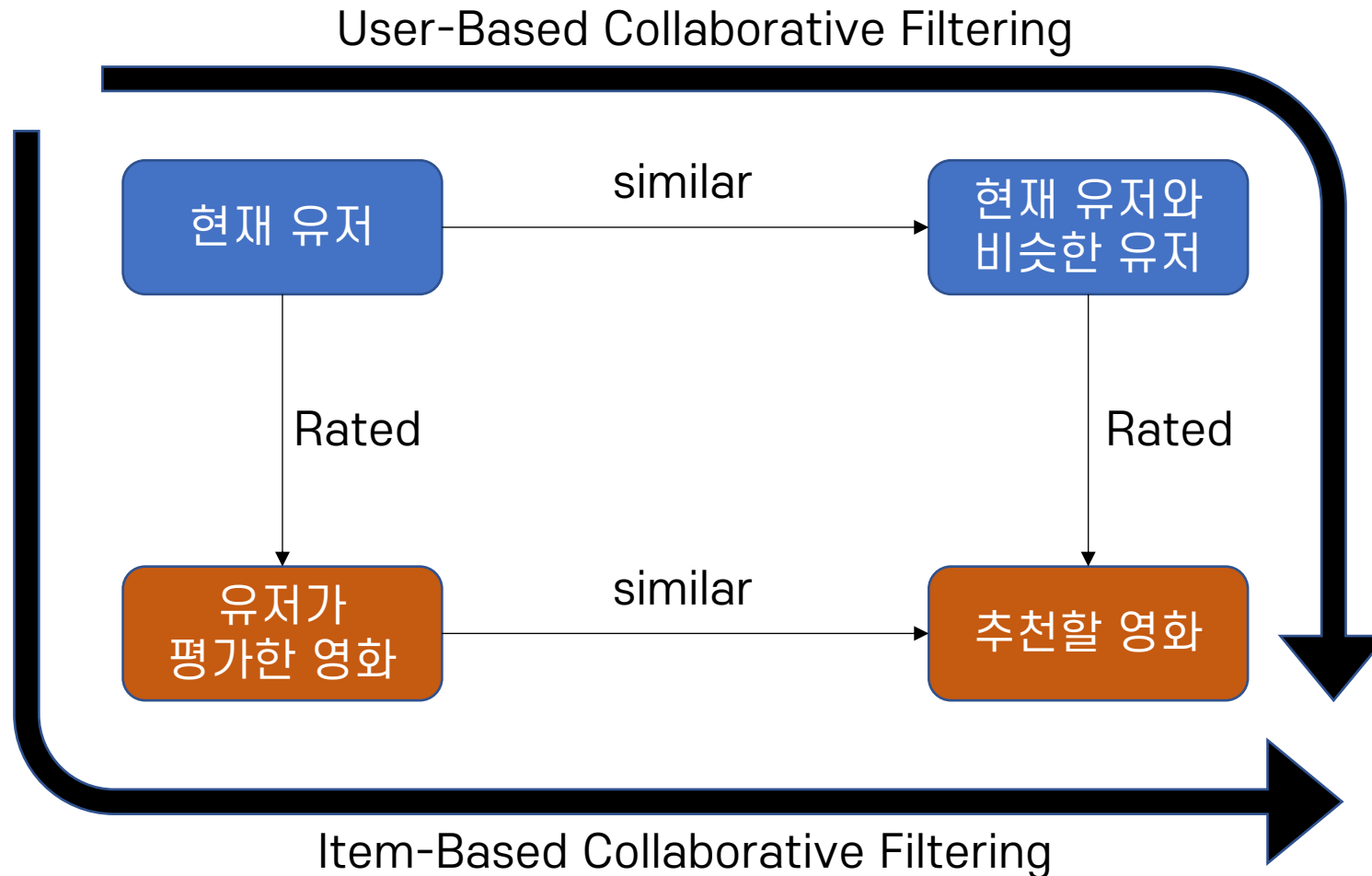
협업 필터링에는 Memory-Based Approach와 Model-Based Approach로 나뉘어짐



Memory Based Collaborative Filtering

Memory-Based 협업 필터링의 종류

Memory-Based 협업 필터링은 크게
User-Based Collaborative Filtering과 Item-Based Collaborative Filtering으로 나뉘어짐














User-Item Matrix

영화 유저	겨울왕국 II	나이비스 이웃	백두산
			
			
			
			









Item Similarity Matrix

	겨울왕국 II	나이비스 이웃	백두산
겨울왕국 II	1	0.25	0.75
나이비스 이웃	0.25	1	0.5
백두산	0.75	0.5	1

User-Item Matrix

영화 유저			
			
			
			
			

User Similarity Matrix

영화 유저				
	1	0.33	0.66	0.66
	0.33	1	0.66	0.66
	0.66	0.66	1	1
	0.66	0.66	1	1

Item-User Matrix

영화 \ 유저				
영화				

행렬 곱

User-Item Matrix

영화 \ 유저				
영화				

=

Similarity Matrix

	1	0.25	0.75
	0.25	1	0.5
	0.75	0.5	1

협업 필터링을 통한 선호도 예측

Item Based Collaborative Filtering으로 영화 추천하기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

예시) 6번 고객은 어떤 영화를 추천할까?

Item Based Collaborative Filtering으로 영화 추천하기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

예시) 6번 고객은 어떤 영화를 추천할까?

6번 고객이 보지 않은 영화 : Heat, GoldenEye

보지 않은 영화 중에서 고객이 선호할만한 영화를 찾자

Item Based Collaborative Filtering으로 영화 추천하기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

예시) 6번 고객은 어떤 영화를 추천할까?

6번 고객이 보지 않은 영화 : Heat, GoldenEye

보지 않은 영화 중에서 고객이 선호할만한 영화를 찾자

6번 고객이 보지 않은 영화에 대한 Rating을 예측하자

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

(1) Item Similarity Matrix 구성하기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

(2) 보지 않은 영화와의 유사도 가져오기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

가정

유사도가 높은 영화에게 유저가 내린 평점과 비슷할 것

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

(2) 보지 않은 영화와의 유사도 가져오기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

해당 영화와 유사도가 높은 K개를 뽑은 후,
평점의 평균을 계산하자
(유사도 기반 가중 평균으로)

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

6번 고객의 영화 HEAT에 대한 평점 값

(2) 보지 않은 영화와의 유사도 가져오기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

해당 영화와 유사도가 높은 K개를 뽑은 후,
평점의 평균을 계산하자
(유사도 기반 가중 평균으로)

평점_{toystory} * 유사도_{toystory} + 평점_{grumpier} * 유사도_{grumpier} + 평점_{sabrina} * 유사도_{sabrina}

유사도_{toystory} + 유사도_{grumpier} + 유사도_{sabrina}

=

5.0 * 0.33 + 3.0 * 0.46 + 5.0 * 0.0

0.33 + 0.46 + 0.0

= 3.85

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

6번 고객의 영화 GoldenEye에 대한 평점 값

$$\frac{\text{평점}_{\text{toystory}} * \text{유사도}_{\text{toystory}} + \text{평점}_{\text{grumpier}} * \text{유사도}_{\text{grumpier}} + \text{평점}_{\text{sabrina}} * \text{유사도}_{\text{sabrina}}}{\text{유사도}_{\text{toystory}} + \text{유사도}_{\text{grumpier}} + \text{유사도}_{\text{sabrina}}} = \frac{5.0 * 0.21 + 3.0 * 0.29 + 5.0 * 0.0}{0.21 + 0.29 + 0.0} = 3.84$$

(2) 보지 않은 영화와의 유사도 가져오기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

해당 영화와 유사도가 높은 K개를 뽑은 후,
평점의 평균을 계산하자
(유사도 기반 가중 평균으로)

Item Based Collaborative Filtering으로 영화 추천하기

예시) 6번 고객은 어떤 영화를 좋아할까?

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
user_id					
2	NaN	4.0	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN
4	NaN	NaN	3.0	NaN	4.0
6	5.0	3.0	NaN	5.0	NaN
7	NaN	3.0	NaN	3.0	NaN
8	4.0	5.0	3.0	NaN	2.0
10	4.0	NaN	NaN	NaN	NaN

(2) 보지 않은 영화와의 유사도 가져오기

	Toy Story	Grumpier Old Men	Heat	Sabrina	GoldenEye
Toy Story	1.00	0.53	0.33	0.50	0.21
Grumpier Old Men	0.53	1.00	0.46	0.54	0.29
Heat	0.33	0.46	1.00	0.00	0.95
Sabrina	0.50	0.54	0.00	1.00	0.00
GoldenEye	0.21	0.29	0.95	0.00	1.00

해당 영화와 유사도가 높은 K개를 뽑은 후,
평점의 평균을 계산하자
(유사도 기반 가중 평균으로)

6번 고객의 영화 Heat에 대한 평점 값 = 3.85

6번 고객의 영화 GoldenEye에 대한 평점 값 = 3.84

6번 고객에게는 영화 Heat를 추천

Memory Based의 한계

User의 수와 Item의 수가 늘어난다면?

movie_id	1	2	3	4	5	6	7	8	9	10	...	111921	112138	112290	112556	112852	116797	117511	117590	118696	125916
user_id																					
1	NaN	3.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	3.0	NaN	NaN	NaN	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	3.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	5.0	NaN	3.0	NaN	NaN	NaN	5.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	NaN	NaN	3.0	NaN	NaN	NaN	3.0	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	4.0	NaN	5.0	NaN	NaN	3.0	NaN	NaN	NaN	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	4.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	4.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.5	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

보지 않은 영화에 대해 모두 가중 평균을 계산해야 함



너무 많은 연산이 필요...