

# DATA MINING

collaborative filtering

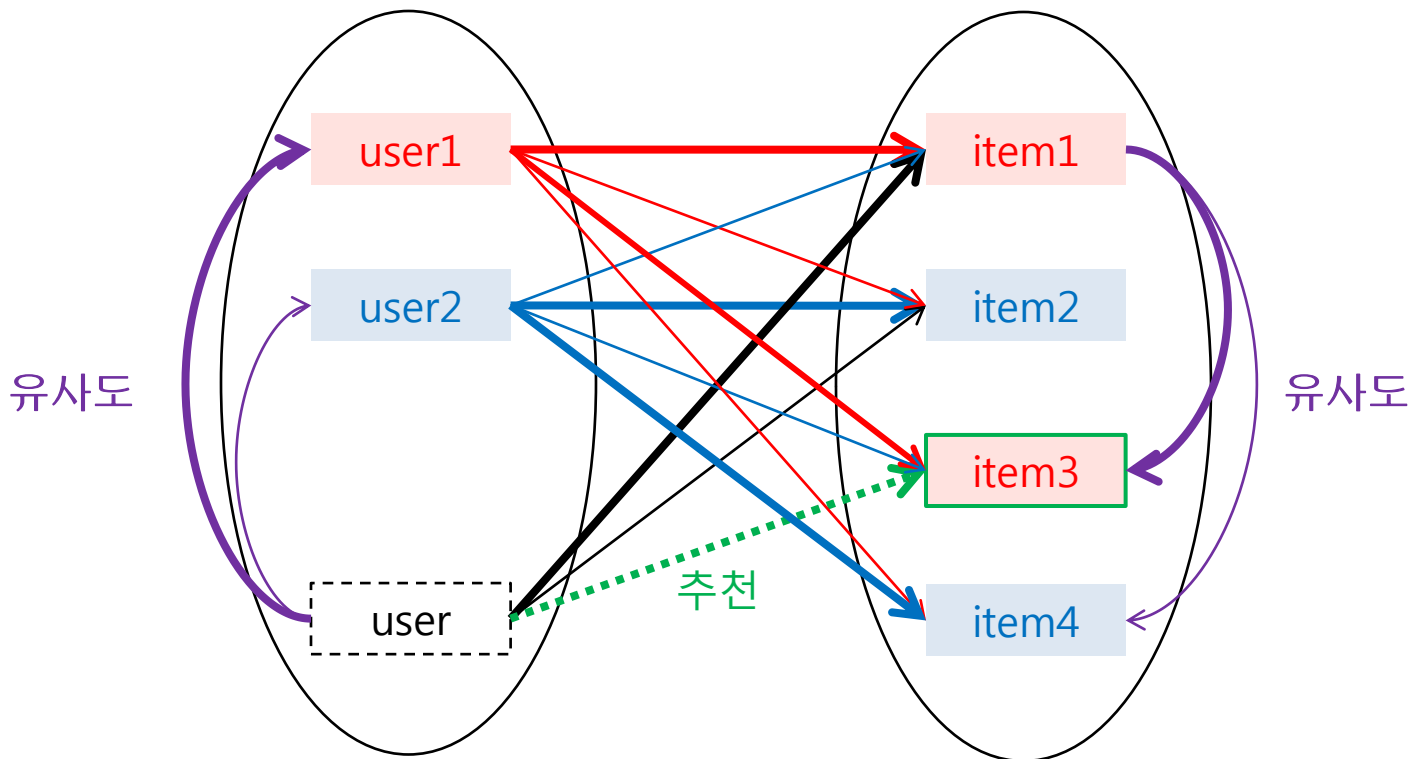
협업 필터링

박혜웅

본 문서는 주관적인 견해가 반영되어 있으므로  
정확한 내용은 관련서적을 참고하시기 바랍니다

# 추천(RECOMMENDATION) 시스템

- user에게 item3, item4중 하나를 추천할 경우 (2가지 방법)
  - user와 user1이 유사하므로 user1이 선호하는 item3를 추천
    - item1, item2에 대한 user와 user1의 선호 성향(유사도)이 비슷
  - user가 item1을 선호하고, item1이 item3와 유사하므로 item3를 추천
    - item1을 선호하는 user1은 item3도 선호함



# 협업필터링(COLLABORATIVE FILTERING)

- ◉ 협업필터링이란

- 큰 집합의 사람들을 비교하여 유사한 취향의 작은 집합을 발견

- ◉ 협업필터링의 사용 예

- 제품 추천시스템 (아마존)

- ◉ 협업필터링의 종류

- 사용자 기반 필터링

- 데이터양이 작고, 데이터 변경이 자주 일어나는 경우
  - 예: 한 반의 모든 학생에 대한 설문조사
- 실시간으로 유사도를 계산

- 항목(제품) 기반 필터링

- 데이터양이 크고, 데이터 변경이 자주 일어나지 않는 경우
  - 예: 대형 사이트에서의 책,영화등 에 대한 추천
- 항목간 유사도를 저장하여 사용.

# 사용자(USER) 기반 협업필터링

## ◉ 사용자(User)기반 협업필터링의 과정

- ① 선호도(평가점수) 조사 및 수치화
  - 사용자별 항목(제품)에 대한 선호도 조사
- ② 유사도 계산
  - 사람간 유사도 계산
- ③ 추천
  - 가장 유사한 사람이 가장 선호하는 항목 추천
    - 비합리적이어서 실제로 사용 안함.
  - 추천가능한 항목에 대하여 모든 사용자의 선호도합을 정규화
    - 그림 참고

# 사용자(USER) 기반 협업필터링

사용자집합	{user1, user2, user3}	preference score(선호도 )	0 ~ 5
나에게 추천가능한 항목집합	{item1, item2}	similarity score(유사도)	0.0 ~ 1.0

	가중치	기본값	계산값	기본값	계산값
	나와의유사도	item1 선호도	(item1 선호도 *나와의유사도)	item2 선호도	(item2 선호도 *나와의유사도)
user1	0.9	5	4.5	1	0.9
user2	0.4			3	1.2
user3	0.1	1	0.1	5	0.5
(유사도*선호도)합	A		4.6		2.9
(유사도)합	B		1.0		1.4
선호도예측값	A/B		4.6		2.0

기본값\*가중치

기본값\*가중치

선호도예측값은 가중평균 을 이용하여 계산하여

- 선호도예측값 = (기본값\*가중치)합 / (가중치) 합 = (유사도\*선호도)합 / (유사도)합 이고,
- 유사도,선호도는 이미 계산되어 있는 값이며, 선호도는 기본값, 유사도는 가중치이다.
- (유사도)합으로 나누는 것은 정규화하는 것이다.

위 표에서 item1의 선호도예측값이 item2보다 높으므로,

- 나에게 추천할 항목은 item1이다.
- 이 때, 가장 큰 영향을 준 것은 나와 가장 유사한 사용자인 user1의 item1에 대한 선호도이다.

# 항목(ITEM) 기반 협업필터링


## ◉ 항목(제품,Item)기반 협업필터링의 과정

- ① 선호도 조사 및 수치화
  - 항목별 사용자의 선호도 조사
- ② 유사도 계산
  - 항목간 유사도 계산
  - 항목간 유사도 정보 저장
    - 항목간 유사도는 사람간 유사도에 비해 자주 변하지 않는다.
- ③ 추천
  - 추천가능한 항목중 가장 유사한 항목을 추천
    - 그림 참고

# 항목(ITEM) 기반 협업필터링

내가 평가한 항목집합	{my1, my2}	preference score(선호도 )	0 ~ 5
나에게 추천가능한 항목집합	{item1, item2, item3}	similarity score(유사도)	0.0 ~ 1.0

	기본값	가중치	예측값	가중치	예측값
	나의 선호도	my1 유사도	(나의 선호도* my1 유사도)	my2 유사도	(나의 선호도* my2 유사도)
item1	5	0.9	4.5	0.1	0.5
item2	3	0.5	1.5	0.5	1.5
item3	1	0.1	0.1	0.9	0.9
(유사도*선호도)합	A		6.1		2.9
(유사도)합	B		1.5		1.5
선호도예측값	A/B		4.0		1.9



선호도예측값은 가중평균 을 이용하여 계산하여

- 선호도예측값 = (기본값\*가중치)합 / (가중치) 합 = (유사도\*선호도)합 / (유사도)합 이고,
- 유사도, 선호도는 이미 계산되어 있는 값이며, 선호도는 기본값, 유사도는 가중치이다.
- (유사도)합으로 나누는 것은 정규화하는 것이다.

위 표에서 item1의 선호도예측값이 item2,item3보다 높으므로,

- 나에게 추천할 항목은 item1이다.
- 이 때, 가장 큰 영향을 준 것은 item1과 가장 유사한 항목인 my1의 item1에 대한 유사도이다.

# 협업필터링 관련 공식

## ◎ 유사도 계산

- 유클리디안 거리(Euclidean distance)
- 피어슨 상관계수(Pearson correlation coefficient)
- 자카드 계수(Jaccard coefficient)
- 맨해튼 거리(Manhattan distance)

## ◎ 선호도예측값 계산

- 가중평균(weighted mean)



# 참고문헌

- ◎ Programming Collective Intelligence
  - Toby Segaran, Oreilly, 2007.08