

추천 시스템

추천시스템은 사용자(user)에게 상품(item)을 제안하는 소프트웨어도구이자 기술이다. 어떤 상품을 구매할지, 어떤 음악을 들을지 또는 어떤 영화를 볼지와 같은 다양한 의사 결정과 연관이 있다.

[협업 필터링(Collaborative Filtering)]

협업 필터링은 사용자의 구매패턴이나 평점을 가지고 다른 사람들의 구매패턴, 평점을 통해서 추천을 하는 방법이다. 추가적인 사용자의 개인 정보나 아이템의 정보가 없어도 추천할 수 있다는 것이 큰 장점이며 2006부터 2009년 동안 열린 Netflix Prize Competition에서 우승한 알고리즘으로 유명세를 떨쳤다.

많은 사용자들로부터 얻은 기호 정보에 따라
사용자들의 관심사들을 자동적으로 예측하게 해주는 방법

가장 일반적인 예로, 온라인 쇼핑몰에서 '이 상품을 구매한 사용자가 구매한 다른 상품들'이란 카테고리로 추천 상품을 보여주는 서비스를 들 수 있다.

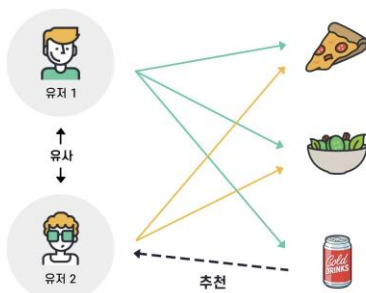
협업 필터링의 포인트는 '많은 사용자들로부터 얻은 기호 정보'이다. 사용자들의 행동이 축적되어감에 따라 추천 결과의 정확도가 더욱 높아지기 때문이다.

협업 필터링의 유형으로는 아래 두 가지가 있다.

1) 사용자 기반 추천 (User-based Recommendation)

나와 비슷한 성향을 지닌 사용자를 기반으로, 그 사람이 구매한 상품을 추천하는 방식이다.

예를 들어 한 사용자가 온라인 물에서 피자과 샐러드, 그리고 콜라를 함께 구매하고, 또 다른 사용자는 피자와 샐러드를 구매했다고 가정해 본다. 알고리즘은 구매 목록이 겹치는 이 둘을 유사하다고 인식하고, 두 번째 사용자에게 콜라를 추천한다. SNS에서의 '친구 추천' 서비스 또한 같은 추천방식이다. 내가 친구로 맺은 사람을 나와 비슷한 성향으로 인식하고 친구의 또 다른 친구들을 나에게도 추천하는 알고리즘이다.

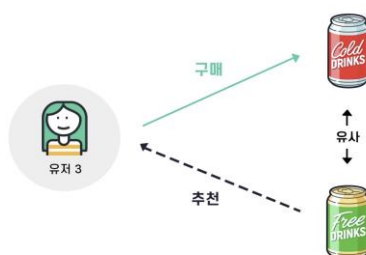


Copyright © 2019 Insignia All rights reserved.

2) 아이템 기반 추천 (Item-based Recommendation)

내가 이전에 구매했던 아이템을 기반으로, 그 상품과 유사한 다른 상품을 추천하는 방식이다.

상품 간 유사도는 함께 구매되는 경우의 빈도를 분석하여 측정한다. 예를 들어 콜라와 사이다가 함께 구매되는 경우가 많다면 콜라를 구매한 사용자에게 사이다를 추천하는 것이다. 아이템 기반 추천은 아래와 같이 도식화할 수 있다.



Copyright © 2019 Insignia All rights reserved.

[협업 필터링 방식의 한계]

두 가지 방식 모두 사용자의 취향을 파악하기 위한 합리적인 방법이다. 그러나 협업 필터링에는 아래와 같은 몇 가지 단점이 있다.

1) 콜드 스타트(Cold Start)

협업 필터링 알고리즘을 사용하기 위해 필수적인 요소는 바로 기존 데이터이다. 사용자 기반 추천방식만으로는 아무런 행동이 기록되지 않은 신규 사용자에게 어떠한 아이템도 추천할 수 없을 것이다. 아이템 기반 추천방식에서도, 새로운 아이템이 출시되더라도 이를 추천할 수 있는 정보가 쌓일 때까지 추천이 어려워진다. 콜드 스타트란 이러한 상황을 일컫는 말이다. '새로 시작할 때의 곤란함' 정도로 해석할 수 있다. 시스템이 아직 충분한 정보를 모으지 못한 사용자에게 대한 추론을 이끌어낼 수 없는 문제이다.

2) 계산 효율 저하

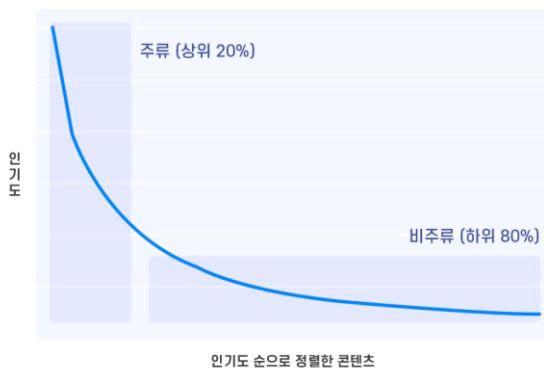
또 다른 문제점은 추천의 효율성이 떨어진다는 점이다. 협업 필터링은 계산량이 비교적 많은 알고리즘이기 때문에, 사용자가 많아질수록 계산이 몇 시간에서 길게는 며칠까지 걸리는 경우도 종종 발생한다. 사용자가 일정 수준 이상이어야 정확한 결과를 낼 수 있는 반면 시간이 더욱 많이 걸리게 된다는 점은 협업 필터링 알고리즘의 문제점이자 딜레마이다.

3) 롱테일(Long Tail)

롱테일은 파레토 법칙(전체 결과의 80%가 전체 원인의 20%에서 일어나는 현상)을 그래프로 나타내었을 때 꼬리처럼 긴 부분을 형성하는 80%의 부분을 일컫는 말이다. 이 현상을 협업 필터링에 적용하면, 사용자들이 관심을 많이 보이는 소수의 콘텐츠가 전체 추천 콘텐츠로 보이는 비율이 높은 '비대칭적 쏠림 현상'이 발생한다는 의미이다. 아이템이나 콘텐츠의 수가 많다고 하더라도 사용자들은 소수의 인기있는 항목에만 관심을 보이기 마련이다. 따라서 사용자들의 관심이 저조한 항목은 정보가 부족하여 추천되지 못하는 문제점이 발생할 수 있다.

미디어 시장의 롱테일을 그래프로 나타내면 다음과 같습니다.

미디어 시장의 Long Tail



협업 필터링의 한계 극복을 위해 '콘텐츠 기반 필터링' 방식을 고안했다.

[콘텐츠 기반 필터링(Contents-based Filtering)]

콘텐츠 기반 필터링은 말 그대로 콘텐츠에 대한 분석을 기반으로 추천하는 방식이다. 영화 콘텐츠의 경우라면 스토리나 등장인물을, 상품이라면 상세 페이지의 상품 설명을 분석한다. 콘텐츠 기반 필터링의 장점은 많은 양의 사용자 행동 정보가 필요하지 않아 콜드 스타트 문제점이 없다는 것이다. 아이템과 사용자 간의 행동을 분석하는 것이 아니라 콘텐츠 자체를 분석하기 때문이다.

넷플릭스에 올라오는 콘텐츠는 50명의 태거(Tagger)에 의해 분류된다. 이들의 역할은 콘텐츠를 면밀하게 보고 여기에 태그를 다는 것입니다. 태거 덕분에 넷플

릭스는 사용자에게 더욱 정교한 추천을 할 수 있다. 사람이 단 태그를 바탕으로 콘텐츠를 5만 종으로 나뉘 정리하기 때문이다. 인터넷 신문사에서는 주로 이 업무를 기계가 실행하며 이때 사용되는 기술이 **텍스트 마이닝** 기술이다.

이 방법에도 문제점이 존재한다. 가장 큰 문제로 꼽히는 것은 '메타 정보의 한정성'이며 상품의 프로파일을 모두 함축하는 데에 한계가 있고 정밀성이 떨어지는 문제가 발생한다.

[최신 알고리즘]













하이브리드 추천 시스템 (Hybrid Recommender Systems)





하이브리드 추천 시스템은 협업 필터링과 콘텐츠 기반 필터링을 조합하여 상호 보완적으로 개발된 알고리즘이다. 협업 필터링의 콜드 스타트 문제 해결을 위해 신규 콘텐츠는 콘텐츠 기반 필터링 기술로 분석하여 추천하고, 충분한 데이터가 쌓인 후부터는 협업 필터링으로 추천의 정확성을 높이는 방식이다.



넘쳐나는 콘텐츠 속에서 살아남기 위한 방법은 정교한 추천 시스템으로 사용자의 리텐션을 늘리고 지속적인 소비를 유도하는 것이다. 유튜브는 추천 시스템의 도입으로 총 비디오 시청 시간을 20 배 이상 증가시켰고, 넷플릭스 자체 평가로는 매출의 75%가 추천 시스템에 의해 발생한다고 한다. 추천이 수익이 되는 시대라고 해도 과언이 아닌 만큼, 많은 개발자들과 연구원들로부터 더욱 발전된 방법들이 개발될 것이라 기대된다.



[사용자 기반 협업 필터링의 예]






		Antman	Avengers	Spiderman	Titanic	Gatsby
user1						
user2						
user3						
user4						

		Antman	Avengers	Spiderman	Titanic	Gatsby
user1		1	1	1	0	0
user2		0	0	0	1	1
user3		1	1	0	0	0
user4		0	0	0	0	1

		Antman	Avengers	Spiderman	Titanic	Gatsby
user1		1	1	1	0	0
		X	X	X	X	X
user3		1	1	0	0	0
		1 + 1 + 0 + 0 + 0 = 2				

		Antman	Avengers	Spiderman	Titanic	Gatsby	Cosine Similarity with user3
user1		1	1	1	0	0	0.81
user3		1	1	0	0	0	1.00

		Antman	Avengers	Spiderman	Titanic	Gatsby	Cosine Similarity with user3
user1		1	1	1	0	0	0.81
user3		1	1	0	0	0	1.00

		Antman	Avengers	Spiderman	Titanic	Gatsby
user1		1	1	1	0	0
user3		1	1		0	0

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

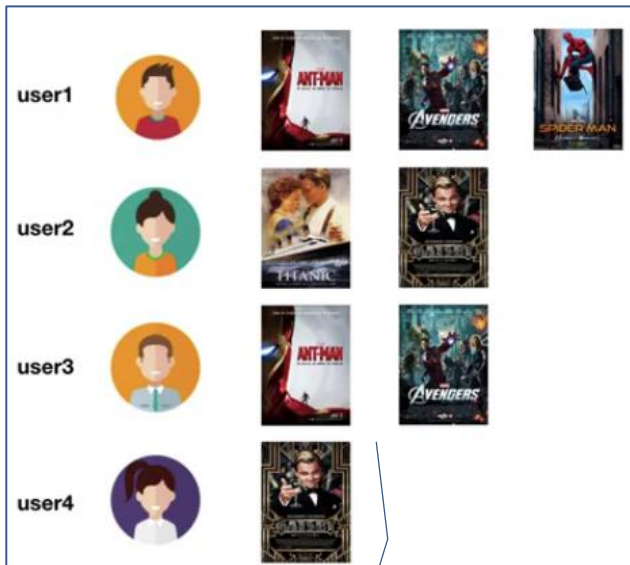
유클리디언 노름 (Euclidean Norm) / 유클리드 거리 함수 / L2 노름

$$\|x\| = \sqrt{x \cdot x} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}$$

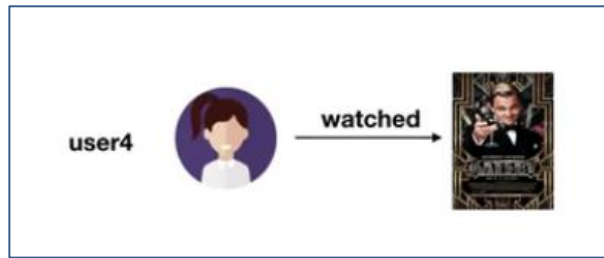
```
import numpy as np
2 / (np.sqrt(3) * np.sqrt(2)) → 0.81
```

```
from numpy.linalg import norm
import numpy as np
def cos_sim(a, b):
    return a @ b / (norm(a)*norm(b))
a = [1,1,1,0,0]
b = [1,1,0,0,0]
cos_sim(np.array(a), np.array(b)) → 0.81
```

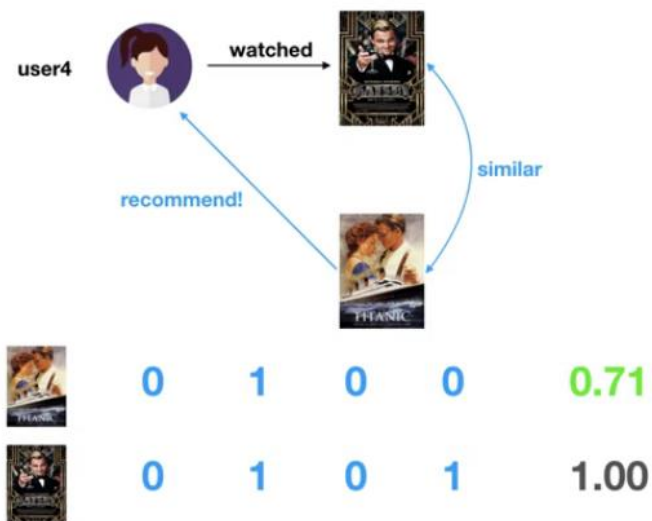
[아이템 기반 협업 필터링의 예]



아이템간의 유사도 분석



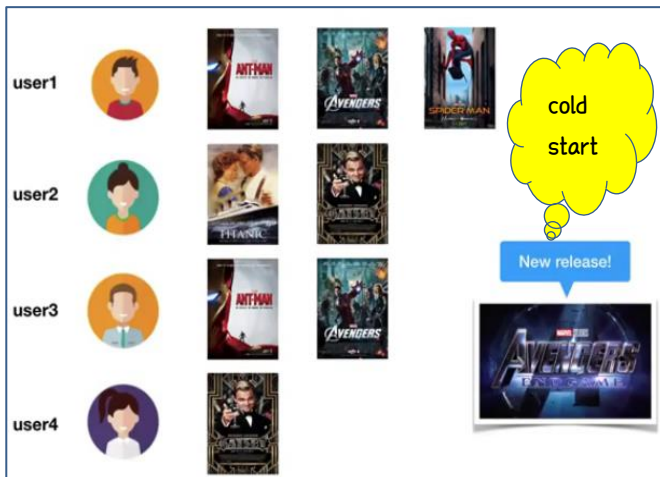
	user1	user2	user3	user4	Cosine Similarity with Gatsby
Ant-Man	1	0	1	0	0
Avengers	1	0	1	0	0
Spider-Man	1	0	0	0	0
Titanic	0	1	0	0	0.71
Gatsby	0	1	0	1	1.00



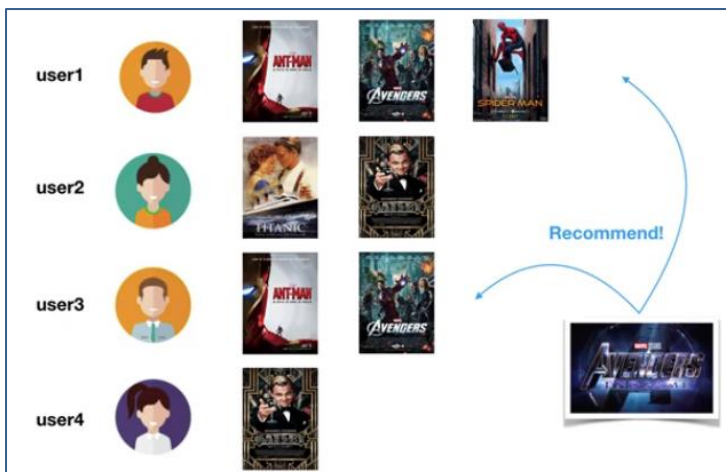
Titanic	0	1	0	0	0.71
Gatsby	0	1	0	1	1.00

[콘텐츠 기반 필터링]

콘텐츠기반 추천시스템은 사용자가 이전에 구매한 상품중에서 좋아하는 상품들과 유사한 상품들을 추천하는 방법이다.



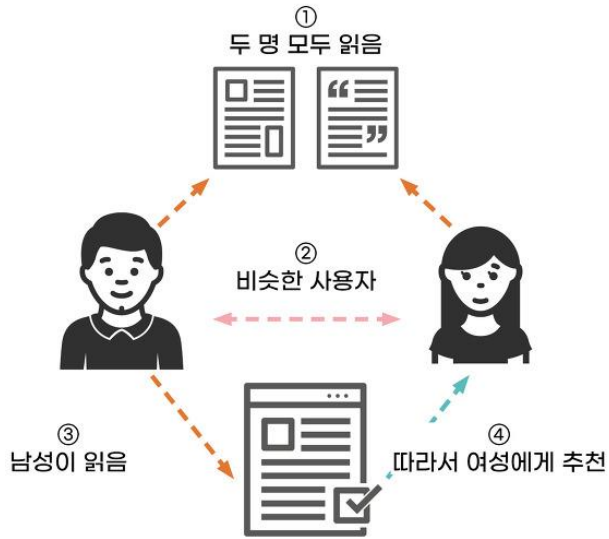
	Iron_Man	Captain_America	Spider_Man	Leonardo_Dicaprio	based_on_true_story
Ant-Man	1	1	1	0	0
The Incredibles	1	1	0	0	0
Spider-Man	1	0	1	0	0
The Incredibles 2	0	0	0	1	1
Avengers: Endgame	0	0	0	1	0



	로맨스	스릴러	액션	공상과학	미스터리	코미디	판타지	범죄
A	1	0	0	0	0	0	1	0
B	0	1	0	0	0	0	1	0
C	0	1	1	0	0	0	0	0
D	0	0	0	1	0	0	1	0
E	0	0	0	0	1	0	0	1
F	1	0	0	0	0	1	0	0

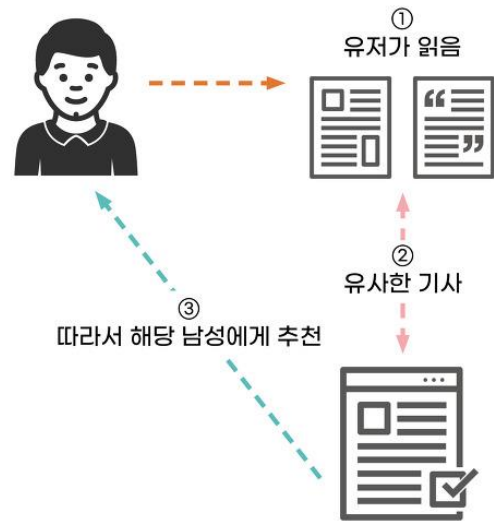
협업 필터링

(Collaborative Filtering)



내용 기반 필터링

(Content-based Filtering)



콘텐츠 기반 필터링

- 사용자가 선호한 콘텐츠를 분석해 추천
- 콘텐츠 자체 정보를 바탕으로 유사도 분석
- 사용자 정보없이 보유한 데이터 만으로 추천 가능
- 필요한 정보는 적지만 범위에 있어 훨씬 제한적임



협업 필터링

- 관심사가 비슷한 다른 사용자들의 행동을 통해 사용자가 관심있어 할만한 항목 추천
 - 많은 양의 데이터가 필요해 데이터가 부족한 초기에 사용이 어려움
- 유저기반 협업 필터링, 아이템기반 협업 필터링

