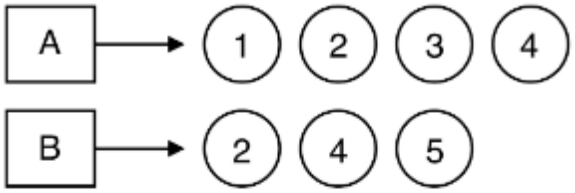
[카카오AI리포트] 내 손안의 AI 비서 추천 알고리즘
https://brunch.co.kr/@kakao-it/72

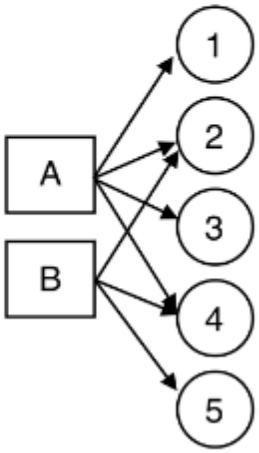
협업 필터링(Collaborative Filtering) – 콘텐츠 사용 패턴이 비슷한 사람들이 서로 비슷한 선호를 가지고 있다고 가정하고 추천을 한다.



a. 스트림 데이터

	A	B
1	1	0
2	1	1
3	1	0
4	1	1
5	0	1

b. 행렬 데이터



c. 그래프 데이터

가장 널리 쓰이는 CF 기술은 행렬분해(Matrix Factorization)기법이다.

MF는 사람들의 기호 데이터를 그림 2의 B와 같은 행렬로 만든다. 모든 사용자가 모든 콘텐츠를 소비할 수는 없기 때문에 이 행렬은 비어있는(sparse) 행렬이 된다.

MF는 이렇게 표현한 행렬의 비어있는 부분을 채우는 방법이다. 토로스(인공지능 추천 플랫폼)은 ALS(Alternating Least Square), BPR-MF(Bayesian Personalized Ranking MF), LMF(Logistic Matrix Factorizaion MF) 등의 기술을 사용하고 있다.

CF를 실제 서비스에 적용할 때는 규모성(scalability)을 고려해야 한다. 사용자 수가 수백만 명이나 되고 새로 등록되는 기사의 개수가 하루에 수십만 개가 넘는다. 큰 데이터가 빠른 속도로 변경되고 있기에 분산 처리나 증분 처리 그리고 알고리즘 최적화가 반드시 필요하다.

CF의 장점 : 클릭 등의 사용자 선호 데이터를 직접 사용하기 때문에 대중적이고, 친숙한 결과를 얻을 수 있다. 사용자 데이터를 직접 최적화하기 때문에 다른 추천 방법론들과 비교하였을 때 성능이 우수하다.

CF의 단점 : 기존 데이터에서 관측하지 못한 콘텐츠를 추천하는 것이 불가능하다.(Cold start Problem) 콘텐츠 소비 이력만을 사용하기 때문에 매우 상이한 콘텐츠도 추천 결과에 노출될 위험이 있으며 인기있는 콘텐츠만 추천 결과에 너무 자주 노출되는 인기 편향 문제가 발생하기도 한다.

콘텐츠 기반 필터링(Contents Based Filtering) – 콘텐츠 자체의 내용을 분석해 유사한 콘텐츠를 찾는 방법론

추천할 콘텐츠들은 글, 사진, 음악이 될 수 있기에 여러 종류의 CB기술이 필요하다.

음악 – 음악의 파형을 직접 분석할 수 있는 신호처리 기술(Spectogram, MFCC)
기사 – 자연언어처리 기술을 통해 글 내용을 분석(word2vec, Latent Dirichlet Allocation)
이미지 – Convolution neural network

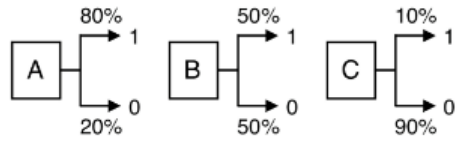
CB의 장점 : 사용자들의 선호 데이터가 없는 콘텐츠도 추천할 수 있고, 내용을 보고 추천을 하기 때문에 생뚱 맞은 콘텐츠가 덜 추천된다는 장점이 있다.

앙상블 기법(Ensemble Method) – 모델별 추천 결과를 조합
대표적인 방법은 2가지
1. 모델마다 가중치를 주어 모델 별 결과를 가중합(weighted sum)
- 상황에 따라 모델별 중요도가 바뀌기 때문에 가중치를 잘 설정하는 것이 중요하다
모델 가중치는 사람이 직접 설정해야 하는 매개변수(hyper-parameter), 하이퍼 파라미터를 최적화 하기 위한 연구가 진행되고 있다..
- 모델 별로 점수가 동일한 의미를 가지는 것이 아니기에 가중합이 잘 통하지 않는 경우가 있다.

2. 모델별 추천 순위를 사용
- 순위 결합(rank aggregation), 상호 랭킹 결합(reciprocal rank fusion), comb mnz
앙상블 방법론들은 사용자 반응을 고려하지 않으므로 모델에 사용자 반응이 반영되기 전까지 사용자 반응에 둔감하다. 사용자들에게 추천 결과가 정적으로 느껴질 수 있다. 반응률이 좋은 콘텐츠를 추천 결과에 더 자주 노출하거나 반응이 나쁜 콘텐츠는 추천 결과에서 빼버리는 등의 매커니즘을 적용하기 어렵다.

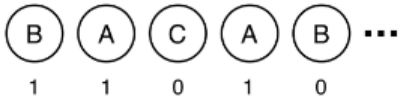
멀티암드밴딧(Multi-armed Bandit)- 서로 다른 승률을 가지고 있는 여러 슬롯 머신에 정해진 횟수만큼 슬롯머신을 시도해볼 수 있을 때, 어떤 순서로 슬롯머신을 시도해야 가장 돈을 많이 벌 수 있는지 찾는 문제

아래와 같이 A, B, C 세 슬롯머신이 있다고 가정했을 때 (a) 이익을 최대화하는 최적해는 (b) 처럼 가장 승률이 좋은 A를 고르는 것이다. 실제로는 A, B, C의 승률을 모르므로 (c) 처럼 다양한 슬롯 머신들을 번갈아가며 승률을 유추해 나가야 좋은 결과를 얻을 수 있다.



a. 슬롯 머신의 예시

b. 예제 문제의 최적해



c. 최적해를 모를 때의 예시 전략

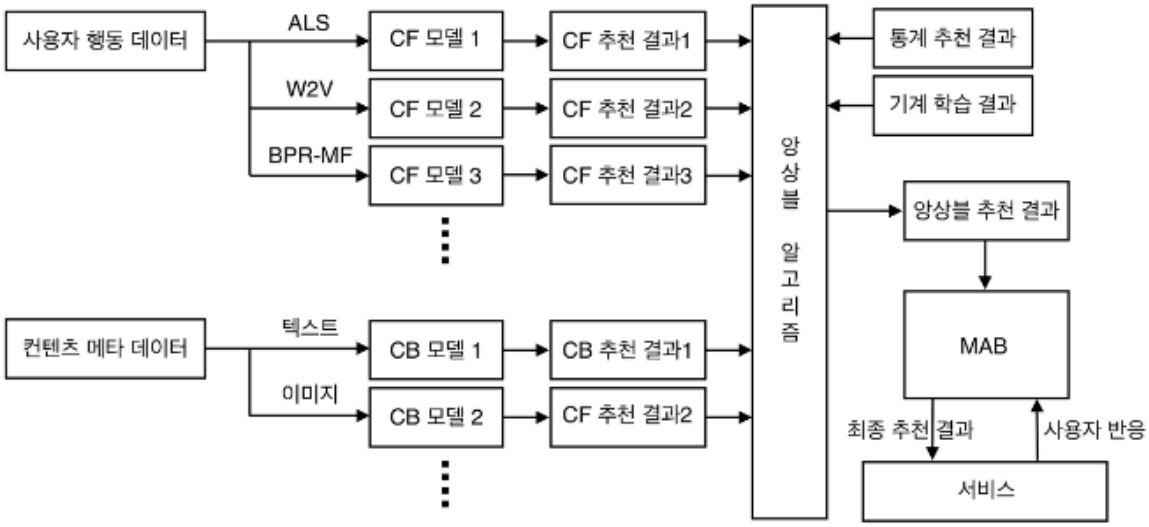
1.임의로 아무 슬롯머신을 시도해보며 슬롯머신들의 승률을 유추한다.(explore)
2.승률이 제일 좋았던 슬롯머신을 시도하는 것을 활용

콘텐츠 추천에 적용을 하면
슬롯머신 = 콘텐츠
시도 = 콘텐츠 추천 결과를 노출
보상 = 클릭
승률 = 반응률

풀어야 할 문제가 많아 이를 바로 사용할 수는 없다.
슬롯머신들의 승률이 한 번 정해지면 같은 승률을 가진다고 가정하는데, 실제로는 콘텐츠 수명이나 트렌드 때문에 승률이 계속 변한다. 위치에 따라 편향이 발생한다.

A/B 테스트 – 사용자들에게 동시에 A와 B결과를 노출하고 이 둘의 결과를 비교하는 지표 측정 방법론이다.

토로스는 CF, CB, 통계 모델, 일반적인 기계학습 모델 등 다양한 모델들에서 추천 결과를 뽑고 뽑은 추천 결과를 앙상블하여 하나의 추천 결과로 병합한다. 만들어진 추천 결과가 사용자들에게 노출되기 시작하면 MAB를 사용해 가장 좋은 추천 결과가 무엇일지 찾아낸다.



[그림 4] 토로스 추천 개요