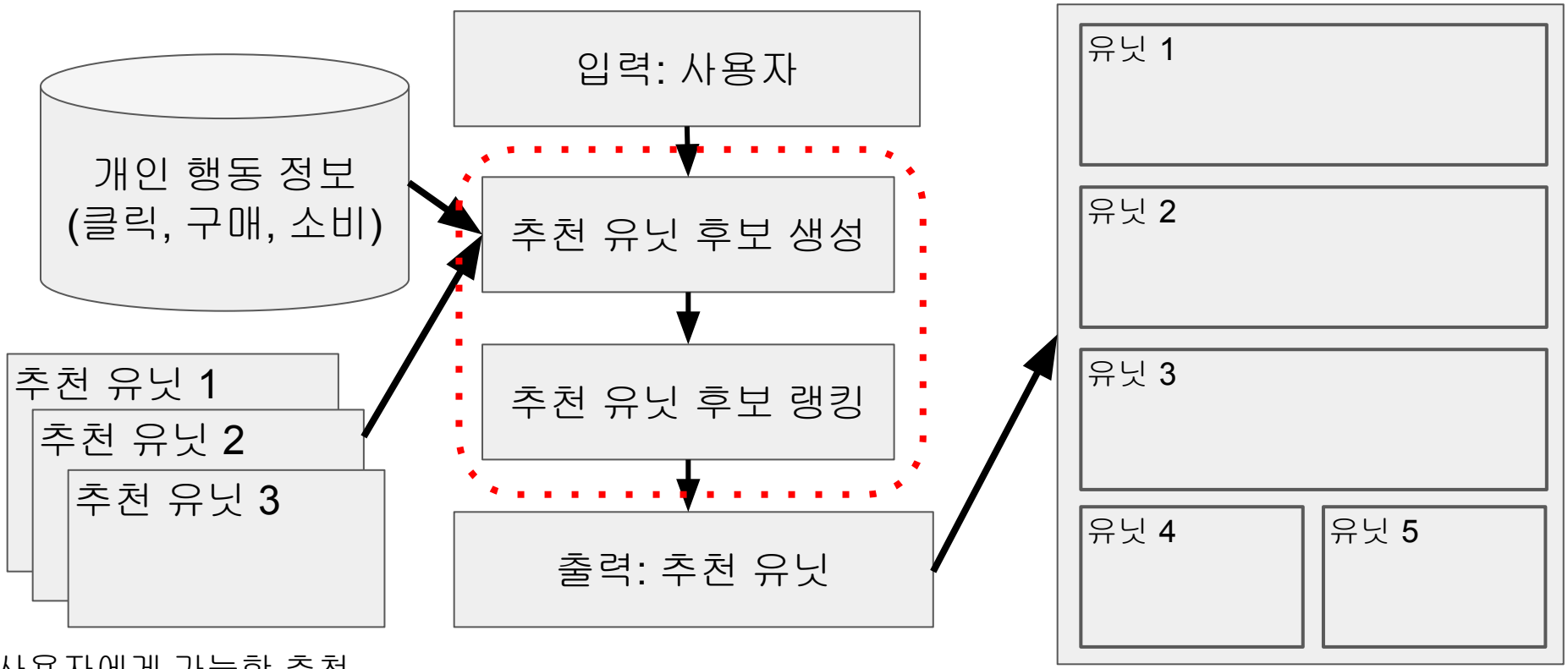
일반적인 추천 엔진 아키텍처

일반적인 추천 엔진 아키텍처에 대해 살펴보자

추천엔진의 기본적인 구조 (전체 추천 페이지 레벨)

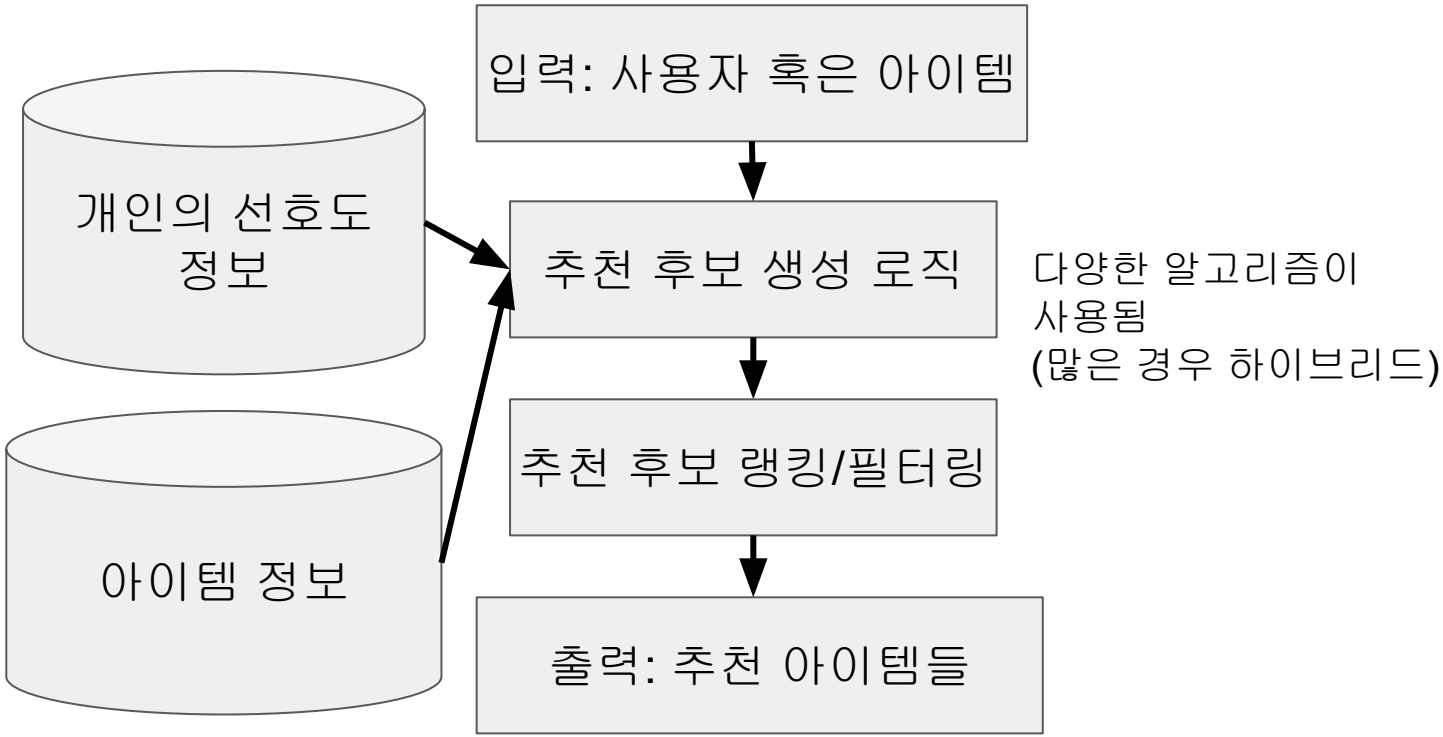


사용자에게 가능한 추천
유닛들

추천엔진의 기본적인 구조 (추천 유닛 레벨)

협업 필터링
평점 예측
모델

콘텐츠 기반
추천



협업 필터링 소개

앞서 알아본 협업 필터링에 대해 더 자세히 알아보자

협업 필터링 (Collaborative Filtering) 소개 (1)

- 기본적으로 다른 사용자들의 정보를 이용하여 내 취향을 예측하는 방식
  a. 사용자 기반 (user to user)
     - 나와 비슷한 평점 패턴을 보이는 사람들을 찾아서 그 사람들이 높게 평가한 아이템 추천
     - “당신과 비슷한 사용자들이 좋아하는 아이템들을 추천합니다"
  b. 아이템 기반 (item to item)
     - 평점의 패턴이 비슷한 아이템들을 찾아서 그걸 추천하는 방식
     - “이 아이템을 좋아한 다른 사용자들이 좋아한 아이템들을 추천합니다"
  c. 예측 모델 기반
     - 앞의 두 개는 기본적으로 유사도를 기반으로 추천 아이템을 결정
     - 이 방식은 평점을 예측하는 머신러닝 모델을 만드는 것

협업 필터링 (Collaborative Filtering) 소개 (2)

- 구현하는 방식에는 크게 두 종류가 존재
  - 메모리 기반
    - 코사인 유사도나 피어슨 상관계수 유사도를 사용해 비슷한 사용자 혹은 아이템을 찾음
    - 평점을 예측할 때는 가중치를 사용한 평균을 사용
    - 이해하기 쉽고 설명하기 쉽지만 스케일하기 힘듬 (평점 데이터의 부족)
  - 모델 기반
    - 머신 러닝을 사용해 평점을 예측 (PCA, SVD, Matrix Factorization, 딥러닝 등등)
      - 딥러닝의 경우에는 오토인코더를 사용하여 차원을 축소함
    - 행렬의 차원을 줄임으로써 평점 데이터의 부족 문제를 해결
    - 하지만 어떻게 동작하는지 설명하기 힘듬 (머신 러닝이 갖는 일반적인 문제)

메모리 기반 vs. 모델 기반

- 메모리 기반은 유사도 함수를 기반으로 비슷한 사용자나 아이템을 검색
  - KNN 방식도 여기에 속한다고 볼 수 있음
  - 평점을 예측하는 것이 아니라 유사도를 기반으로 추천
- 모델 기반은 어떤 비용 함수를 기반으로 학습
  - 즉 머신 러닝!
    - 넷플릭스 프라이즈에서는 RMSE를 지표로 사용 (실제 평점과 예측 평점간의 차이)
  - SVD++(Singular Vector Decomposition++)에서는 SGD (Stochastic Gradient Descent)를 사용하여 (딥러닝처럼) 학습
  - 딥러닝에서는 오토인코더를 사용하여 사용자 아이템 행렬의 패턴을 배움
    - 오토인코더란?
      - 딥러닝에서 데이터 차원을 축소하는 방식으로 인코딩을 통해 데이터를

협업 필터링 (혹은 일반적인 추천 엔진) 평가

- 메모리 기반 협업 필터링
  - 평점의 예측 없이 유사도 기반으로 추천할 아이템을 결정하는 방식
    - 그러기에 RMSE와 같은 평점 기반 평가는 불가
  - 보통 Top-N(혹은 nDCG) 방식으로 평가
    - 사용자가 좋아한 아이템을 일부 남겨두었다가 추천 리스트에 포함되어 있는지 보는 방식
    - 추천 순서를 고려해서 평가하면 nDCG (normalized Discounted Cumulative Gain)
- 모델 기반 협업 필터링
  - 머신러닝 알고리즘들이 사용하는 일반적인 방식(예: RMSE)으로 성능 평가 가능
  - 메모리 기반에서 사용하는 Top-N이나 nDCG 방식도 사용 가능
- 온라인 테스트 (A/B 테스트)
  - 가장 좋은 방식은 실제 사용자에게 노출시키고 성능을 평가하는 것

SurpriseLib 소개

- 협업 필터링과 관련한 다양한 기능을 제공하는 라이브러리
  - KNNBasic 객체 이용해서 사용자 기반 혹은 아이템 기반 협업 필터링 구현
  - SVD 혹은 SVDpp 객체를 이용해서 모델 기반 협업 필터링
- 협업 필터링 알고리즘의 성능 평가를 위한 방법 제공
- 실습에서 사용해볼 예정
  - scikit-learn 사용법과 비슷

사용자 기반 협업 필터링

사용자의 유사도를 기준으로 추천을 하는 사용자 기반 협업 필터링에 대해 알아보자

사용자의 유사도 측정 (1)

- 사용자들을 벡터(아이템에 대한 평점)로 표현
- 지정된 사용자(u)와 다른 나머지 사용자들과 유사도 측정
  - 공통으로 평점을 준 아이템만 대상으로 사용자들간 유사도 측정
    - u와 다른 사용자들 하나씩 (u') 대상으로 유사도 측정

$$sim(u,u') = cos(\theta) = \frac{\mathbf{r}_u \cdot \mathbf{r}_{u'}}{\|\mathbf{r}_u\| \|\mathbf{r}_{u'}\|} = \sum_i \frac{r_{ui} r_{u'i}}{\sqrt{\sum_i r_{ui}^2} \sqrt{\sum_i r_{u'i}^2}}$$

$$\hat{r}_{ui} = \frac{\sum_{u'} sim(u,u') r_{u'i}}{\sum_{u'} |sim(u,u')|}$$

- 이 중 가장 비슷한 K명의 사용자를 선택 (Top K)
- 이 K명의 사용자들(u') 대상으로 u가 평가하지 않은 아이템을 i
  - u'가 평가한 아이템 i의 평점에 u와 u'의 유사도를 가중치로 합을 계산
  - 합산한 값이 큰 아이템들을 추천

사용자의 유사도 측정 (2)

- 사용자 대 아이템 행렬을 사용자간 유사도 행렬으로 변환

	scifi1	scifi2	scifi3	comedy1	comedy2	comedy3
user1	4.0	5.0	3.0	NaN	2.0	1.0
user2	5.0	3.0	3.0	2.0	2.0	NaN
user3	1.0	NaN	NaN	4.0	5.0	4.0
user4	NaN	2.0	1.0	4.0	NaN	3.0
user5	1.0	NaN	2.0	3.0	3.0	4.0

	user1	user2	user3	user4	user5
user1	1.000000	0.764091	-0.837183	-0.577861	-0.736161
user2	0.764091	1.000000	-0.780626	-0.601483	-0.746852
user3	-0.837183	-0.780626	1.000000	0.233344	0.641732
user4	-0.577861	-0.601483	0.233344	1.000000	0.245501
user5	-0.736161	-0.746852	0.641732	0.245501	1.000000

사용자 기반 협업 필터링 실습

- 더미 데이터를 이용한 사용자 기반 추천 실습:
  - https://colab.research.google.com/drive/14lLkl-ddezoOR7srZWJ2vBAEz1sYsSmY?usp=sharing
- 사용자 기반 협업 필터링 추천:
  - https://colab.research.google.com/drive/1h92tHHxKLk6TKuGb6j-6umQfJEYTD1SH?usp=sharing

아이템 기반 협업 필터링

아이템의 유사도를 기준으로 추천을 하는 사용자 기반 협업 필터링에 대해 알아보자

아이템의 유사도 측정

- 주어진 아이템을 기반으로 가장 비슷한 아이템을 찾아서 추천
  - i가 메인 아이템. j는 비교 대상이 되는 아이템
  - 분자: i, j를 모두 평가한 사용자(u)를 대상으로 i와 j간의 유사도를 계산해서 합산
  - 분모: 모든 사용자(u)의 아이템 i 평점을 제곱 후 합산하여 루트 계산 x 모든 사용자(u)의 아이템 j 평점을 제곱 후 합산 하여 루트 계산
  - 최종적으로 i와 유사도가 가장큰 j를 추천 (N개)

$$similarity(i,j) = \frac{\sum_u^U r_{(u,i)} r_{(u,j)}}{\sqrt{\sum_u^U r_{(u,i)}^2} \sqrt{\sum_u^U r_{(u,j)}^2}}$$

Similarity equation (by Muffaddal)

아이템 i와 j간의 유사도 측정

- 아이템 대 사용자 행렬을 아이템간 유사도 행렬도 변환

	user1	user2	user3	user4	user5
scifi1	4.0	5.0	1.0	NaN	1.0
scifi2	5.0	3.0	NaN	2.0	NaN
scifi3	3.0	3.0	NaN	1.0	2.0
comedy1	NaN	2.0	4.0	4.0	3.0
comedy2	2.0	2.0	5.0	NaN	3.0
comedy3	1.0	NaN	4.0	3.0	4.0

	scifi1	scifi2	scifi3	comedy1	comedy2	comedy3
scifi1	1.000000	0.706689	0.813682	-0.799411	-0.025392	-0.914106
scifi2	0.706689	1.000000	0.723102	-0.845154	-0.518999	-0.843374
scifi3	0.813682	0.723102	1.000000	-0.847946	-0.379980	-0.802181
comedy1	-0.799411	-0.845154	-0.847946	1.000000	0.148039	0.723747
comedy2	-0.025392	-0.518999	-0.379980	0.148039	1.000000	0.393939
comedy3	-0.914106	-0.843374	-0.802181	0.723747	0.393939	1.000000

아이템 기반 협업 필터링 실습

- 더미 데이터를 이용한 아이템 기반 추천 실습:
  - https://colab.research.google.com/drive/1SJT7Q4NHvStl1UOOZlKqo1fCBQ5T4MPy?usp=sharing
- 아이템 기반 협업 필터링 추천:
  - https://colab.research.google.com/drive/1znUtE2MGj_9tJeUCVeQGjX3k21OaJ1mu?usp=sharing