추천시스템 정리

추천시스템 종류

참고 : https://koreascience.kr/article/JAKO201512053817215.pdf

1. 협업 필터링

- 주요 가정

  `많은 사용자로부터 얻은 취향 정보를 토대로 나와 비슷한 취향을 가진 사람들이 선호하는 콘텐츠를 나도 좋아할 가능성이 크다.`

- Matrix Completion (유저-아이템 행렬)의 빈 부분을 채우기 위해 다수가 협업하는 식으로 해결하는 모델

  → 각 유저, 아이템은 특정 수준의 상관관계를 가진다고 가정함.

- ex) A와 B가 유사한 그룹으로 묶인다면 B가 선호하는 아이템을 A가 좋아할 것으로 예측. A가 구매한 아이템을 제외하고 B가 선호하는 아이템을 A에게 추천함.

- **Memory-Based methods**

  - **유사도 측정**

    - 점수 기반

      `피어슨 상관계수`

      `코사인 유사도`

      내적 공간의 두 벡터 간 각도의 코사인 값 이용 → 1이면 같은 성향, 0이면 다른 성향

      but 사용자들이 다른 평가 척도를 사용할 경우 유사 정도 파악하기 어려움

      → `보완 코사인 유사도` 제안 (사용자의 평균 평가점수 대비 특정 아이템에 대한 선호도를 파악할 수 있어 정규화된 유사도 값을 얻을 수 있음)

    - 순위 기반

      `스피어만 순위 상관계수`

      사용자 a와 b의 점수를 각각 순위로 변환한 후 차이를 통해 유사도를 측정함.

▼ 유사도 공식

- 피어슨 상관계수

$$w_{a,b} = \frac{\sum_{i \in I}(r_{a,i} - \overline{r_a})(r_{b,i} - \overline{r_b})}{\sqrt{\sum_{i \in I}(r_{a,i} - \overline{r_a})^2} \sqrt{\sum_{i \in I}(r_{b,i} - \overline{r_b})^2}} \quad (6)$$

- 코사인 유사도

$$w = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\| \vec{i} \| * \| \vec{j} \|} \quad (8)$$

- 보완 코사인 유사도

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_u})(r_{u,j} - \overline{r_u})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_u})^2}} \quad (9)$$

- 스피어만 순위 상관계수

$$w_{a,b} = 1 - \frac{6 \sum_{i \in I} d_i^2}{n(n^2 - 1)} \quad (10)$$

◦ **선호도 예측**

▪ 가중합

사용자 기반 협업 필터링에서 사용

추천 대상 고객 a가 아이템 i에 대해 갖는 예측 선호도

추천 대상 고객과 사용자 간 유사도가 높을수록 큰 가중치 부여

- 단순가중평균

  아이템 기반 협업 필터링에서 사용

  예측하고자 하는 아이템과 다른 아이템과의 유사도를 가중치로 부여

  유사한 아이템 점수를 보다 크게 반영해 예측값 계산

▼ 예측 계산 공식

- 가중합

$$P_{a,i} = \overline{r_a} + \frac{\sum_{u \in U} (r_{u,i} - \overline{r_u}) \times w_{a,u}}{\sum_{u \in U} |w_{a,u}|} \quad (11)$$

- 단순가중평균

$$P_{a,i} = \frac{\sum_{n \in N} r_{a,n} \times w_{i,n}}{\sum_{n \in N} |w_{i,n}|} \quad (12)$$

- **상위 N개 아이템 추천**
  - 추천 대상 고객의 선호도가 가장 높을 것이라 예상되는 상위 N개의 아이템을 최종적으로 선택해 추천 대상 고객에게 아이템 목록 제공

▼ 추천 알고리즘 예시

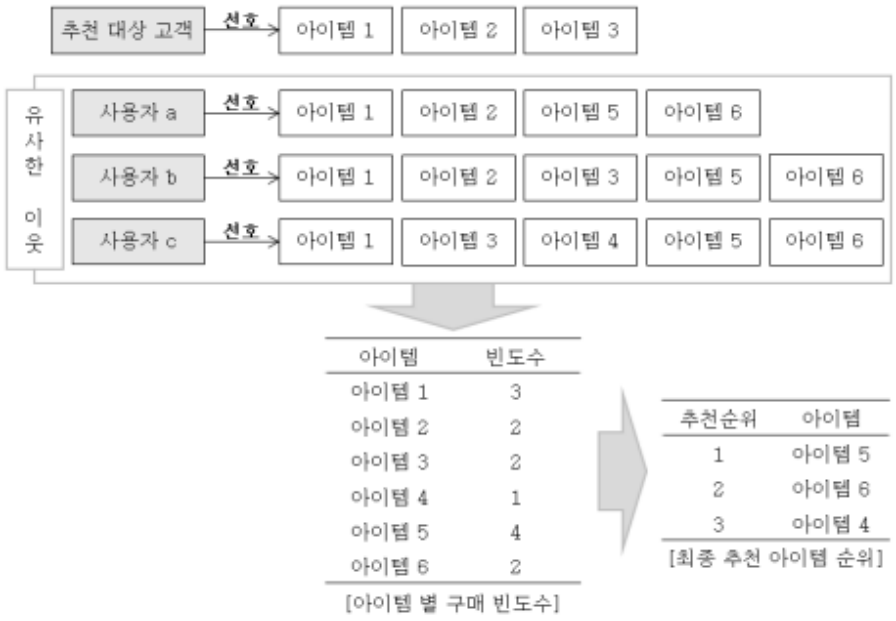

Figure 5. User-based top-n recommendation algorithm

(a) Top-n lists of recommended items

(b) Ranking of recommendation

Figure 6. Item-based top-n recommendation algorithm

- **User-based collaborative filtering**
  - 유저 간의 유사도가 높을수록 높은 가중치를 부여
  - 특정 유저가 구매하지 않았으나 동질 그룹의 다른 유저가 선호하는 아이템을 추천

        - 일반적으로 특정 A와 유사한 Top K의 유사한 유저들과 동질 그룹으로 구성해 A가 선호할만한 아이템을 선정
    - **Item-based collaborative filtering**
        - 아이템 간 유사도가 높을수록 높은 가중치를 부여

          ▼ 유사도

            - 피어슨 상관계수

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2}\sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_j})^2}} \qquad (7)$$

        - B라는 아이템에 대한 A 유저의 선호도를 예측하기 위해, B와 가장 유사한 Top K 아이템을 선정하여 Item set을 구성
- **Model-based methods**
    - 머신러닝이나 데이터마이닝 방법에서 예측 모델의 context를 기반한 방법
    - 모델이 파라미터화되어 있다면, 이 모델의 파라미터는 context 내에서 학습
    - 잠재 요인 협업 필터링
        - 사용자와 아이템 간 평점 행렬 속에 숨어있는 잠재 요인 행렬을 추출해 내적곱을 통해 사용자가 평가하지 않은 항목들에 대한 평점까지 예측해 추천
        - `Matrix Factorization` 방법을 통해 큰 다차원 행렬을 차원 감소시키는 과정에서 행렬에 포함되어 있는 잠재 요인들을 추출할 수 있음

2. 콘텐츠 기반 추천시스템

- 사용자가 과거에 경험했던 아이템 중 비슷한 아이템을 현재 시점에서 추천
- 정보를 찾는 과정과 과거 정보를 활용해서 유저의 성향을 배우는 문제
- 시스템 구조

- 분석 과정
  - 해당 아이템을 설명할 수 있는 특징 벡터화 후 아이템 간 유사도 계산해 높은 아이템 추천
    - 텍스트의 경우 `키워드 분석` 혹은 `의미 분석`으로 유사도 계산함
    - 키워드 분석

      각 텍스트 아이템에서 키워드를 추출한 뒤 아이템 간 키워드를 비교해 유사도 계산

      → `TF-IDF`가 대표적

      ▼ TF-IDF

      - `TF`
        - 한 아이템 내에서 특정 단어가 출현한 빈도수
        - 많이 나타날수록 상대적으로 중요하다는 가정
        - 공식

$$TF(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{z,j}} \quad (2)$$

      - `IDF`
        - 출현 빈도가 높은 불용어를 보완하기 위한 방법
        - 상대적으로 적은 아이템에서 출현한 단어일수록 높은 IDF 값을 가져 키워드로 추출될 확률이 높음
        - 공식

$$IDF(t_k, d_j) = \log \frac{N}{n_k} \quad (3)$$

      - 단어 가중치 (0부터 1 사이)

$$\text{단어 가중치}(t_k, d_j) = \frac{TF-IDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} TF-IDF(t_k, d_j)^2}} \qquad (4)$$

- 단어 가중치가 높은 상위 N개의 단어가 키워드로 선택됨.

  - 유사도 계산 : 유클리디안, 코사인, 피어슨 유사도

    → 여러 가지를 적용 및 비교해 최적의 결과를 이용

- 장/단점

  - 다른 사용자의 데이터가 존재하지 않더라도 신규 사용자에게 콘텐츠 추천 가능
  - 추천하는 콘텐츠에 대한 근거 제시 가능
  - 새로 추가된 콘텐츠나 유명하지 않은 콘텐츠에 대한 추천 가능

    → 현재 필드에서 가장 많이 사용되는 이유

  but 과거 이력에 대한 정보 없을 때 추천 어려움 + 이미 알고 있는 유사한 콘텐츠만 추천하는 문제 발생 가능

3. Knowledge based recommend system

- 사용자의 구매 이력이 적은 경우 사용
- 아이템을 추천하기 전에 아이템의 특징과 명시적인 질문을 통해 획득한 사용자 선호도와 추천 범위 등 아이템들에 대한 정보를 고려해 추천

  → 사용자 profile 사용

4. Hybrid recommend system

- 위의 다양한 추천 시스템을 결합
- 한 서비스의 단점을 다른 서비스의 장점으로 만듦
- ex. 새로운 아이템에 대한 평점이 없으면 추천 성능이 떨어지게 되는 협업 필터링과 아이템의 특징에 대한 정보를 이용할 수 있는 지식 기반 추천 시스템 결합

추천 시스템 사례

1. 넷플릭스

참고 : https://help.netflix.com/ko/node/100639

- 콘텐츠 분류 및 태깅을 크게 두 가지 방법으로 진행
  - 콘텐츠의 `기본 데이터` (출시 연도, 감독, 출연자 등 객관적인 정보)
  - `전문가 데이터` (긴장감 넘치는, 불안한 등 주관적인 정보)

    : 넷플릭스 양자이론 (등록된 콘텐츠를 더 이상 쪼갤 수 없는 수준까지 쪼갠다는 의미)을 기준으로 분석 전문가가 모든 콘텐츠를 감상하고 분석함
- 고객 데이터 중 `사용자 의식 데이터` 사용

  → 좋아요/싫어요 결과와 이것이 없지만 재시청했을 경우 관심 있는 영화로 판단