협업 필터링

이청용 교수

leecy@hansung.ac.kr

강의개요

- **협업 필터링 개념**
  - ✓ 기본 개념
  - ✓ 협업 필터링의 방식
- **협업 필터링의 주요 방식**
  - ✓ 사용자 기반 협업 필터링
  - ✓ 아이템 기반 협업 필터링
- **유사도 측정 방법**
  - ✓ 코사인 유사도 (Cosine Similarity)
  - ✓ 피어슨 상관계수 (Pearson Correlation Coefficient)
  - ✓ 코사인 유사도 vs. 피어슨 상관계수
- **선호도 예측**
  - ✓ 가중평균 (Weighted Average)
  - ✓ 평균 보정 방식 (Mean-Centered Rating)
  - ✓ 가중평균과 평균 보정 방식의 비교
- **성능 평가 지표**
  - ✓ 예측 성능 평가 지표
  - ✓ 순위 기반 평가 지표
- **협업 필터링의 한계점**
  - ✓ 콜드 스타트 문제 (Cold Start Problem)
  - ✓ 데이터 희소성 문제 (Data Sparsity)
  - ✓ 확장성 문제 (Scalability Problem)
  - ✓ 군집화 문제 (필터 버블, Filter Bubble)
  - ✓ 잠재적 편향성 문제 (Bias Problem)

협업 필터링 개념

기본 개념

- **협업 필터링 정의**
  - ✓ 사용자 행동 데이터를 분석하여 유사한 사용자나 아이템 간의 관계를 바탕으로 추천을 제공하는 방식
  - ✓ 사용자가 남긴 **평점, 클릭, 구매 기록** 등 다양한 상호작용 데이터를 활용하여 사용자가 좋아할 가능성이 높은 아이템을 찾아 추천
- **기본 아이디어**
  - ✓ 협업 필터링은 **콘텐츠 자체의 특성**이나 **메타데이터**(예: 영화 장르, 책의 저자 등)를 고려하지 않음
  - ✓ 사용자들이 특정 아이템에 대해 공동으로 참여한 패턴을 기반으로 유사한 취향을 가진 다른 사용자들이 선호한 아이템을 추천하거나, 유사한 아이템을 추천하는 방식
- **사용자-아이템 상호작용 행동 데이터 활용**
  - ✓ **평점 데이터:** 사용자가 특정 아이템에 대해 부여한 점수 (예: 영화 평점 5점 만점에 4점)
  - ✓ **구매 데이터:** 사용자가 특정 아이템을 구매한 기록 (예: 사용자가 책을 구매한 기록)
  - ✓ **클릭 및 탐색 기록:** 사용자가 특정 아이템을 클릭하거나 탐색한 기록
  - ✓ **장바구니 및 위시리스트 데이터:** 사용자가 아이템을 장바구니에 담거나 위시리스트에 추가한 기록

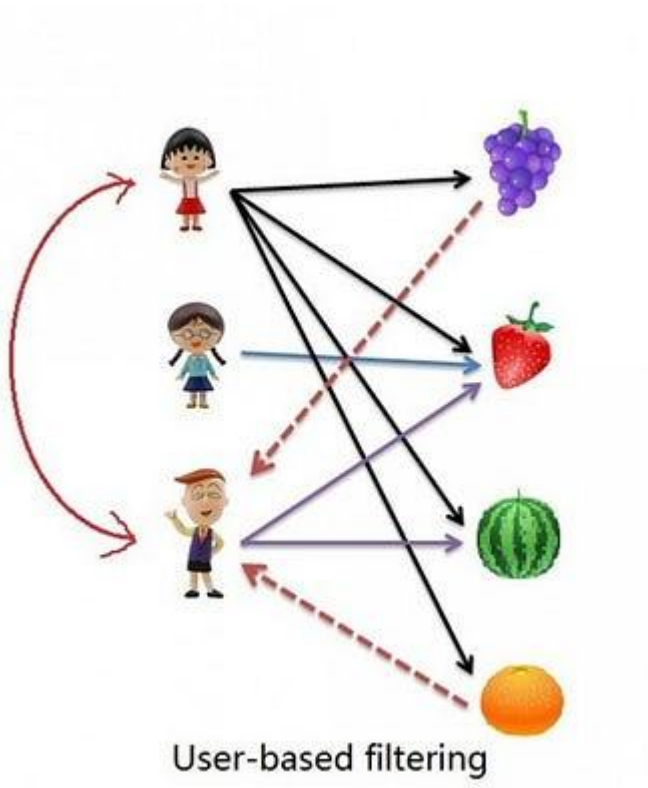
협업 필터링의 방식

- **사용자 기반 협업 필터링(User-Based Collaborative Filtering)**
  - ✓ 유사한 취향을 가진 다른 사용자가 좋아한 아이템을 추천하는 방식
    - 예를 들어, A 사용자와 B 사용자가 유사한 취향을 가지고 있다고 판단되면 B 사용자가 높게 평가한 아이템을 A 사용자에게 추천
  - ✓ 사용자 간의 상호작용을 활용하여 추천을 생성하며 추천 대상 사용자와 비슷한 다른 사용자들의 평점을 기반으로 추천
- **아이템 기반 협업 필터링(Item-Based Collaborative Filtering)**
  - ✓ 사용자가 선호한 아이템과 유사한 다른 아이템을 추천하는 방식
    - 예를 들어, A 사용자가 특정 영화를 좋아하면 해당 영화와 비슷한 영화를 추천
  - ✓ 아이템 간의 유사성을 활용하며 사용자 간의 상호작용 데이터보다는 아이템 자체에 대한 평가 기록을 중요하게 고려

협업 필터링의 주요 방식

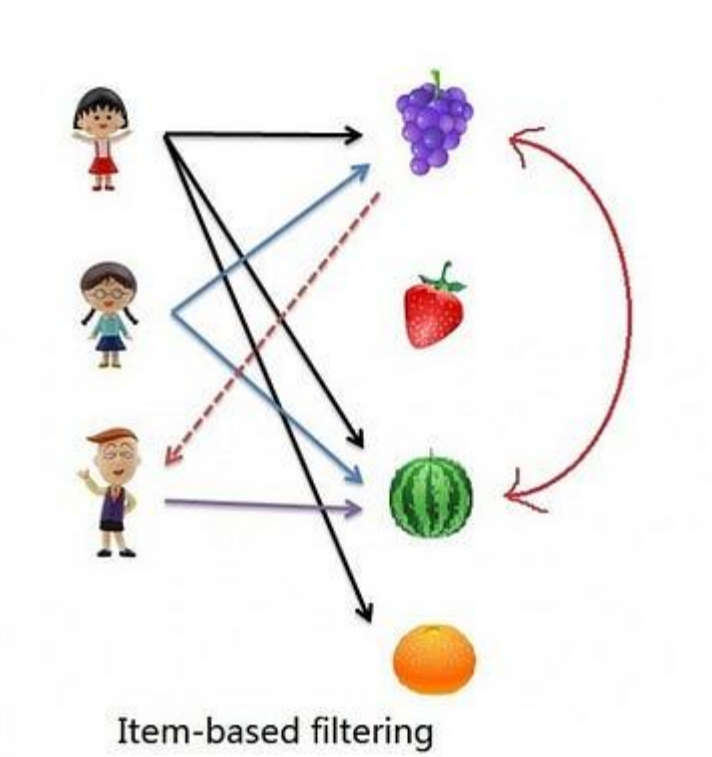
사용자 기반 협업 필터링

- **기본 개념**
  - ✓ 사용자 기반 협업 필터링은 유사한 사용자들의 행동을 활용해 추천을 생성하는 방식
  - ✓ 특정 사용자가 어떤 항목을 선호할지를 예측할 때 비슷한 취향을 가진 사용자들의 행동 데이터를 분석하여 추천을 제공
- **작동 원리**
  - ✓ **유사한 사용자 탐색**
    - 먼저 추천을 받는 사용자와 비슷한 취향을 가진 사용자들을 탐색
    - 다양한 유사도 측정 방법(코사인 유사도, 피어슨 상관계수 등)을 사용하여 사용자 간의 유사도를 계산
  - ✓ **추천 생성**
    - 유사한 사용자들이 선호하는 아이템 중에서 추천 대상 사용자가 아직 평가하지 않은 항목을 찾아 추천



아이템 기반 협업 필터링

- **기본 개념**
  - ✓ 아이템 기반 협업 필터링은 유사한 아이템들 간의 관계를 이용해 추천을 생성하는 방식
  - ✓ 특정 사용자가 선호할 아이템을 예측할 때 해당 사용자가 이미 선호한 아이템들과 유사한 아이템을 찾아 추천을 제공
- **작동 원리**
  - ✓ **유사한 아이템 탐색**
    - 사용자가 선호한 아이템과 유사한 특성을 가진 다른 아이템을 탐색
    - 아이템 간의 유사도는 사용자들이 아이템을 어떻게 평가했는지에 따라 결정
  - ✓ **추천 생성**
    - 사용자가 선호하는 아이템과 유사한 아이템을 추천 리스트를 생성



유사도 측정 방법

코사인 유사도 (Cosine Similarity)

- **정의**
  - ✓ 두 벡터 간의 각도를 기준으로 유사성을 측정하는 방법
  - ✓ 벡터의 크기보다는 방향에 초점을 맞춤
  - ✓ 주로 아이템 기반 협업 필터링에서 사용
- **계산 방식**
  - ✓ 두 벡터의 내적(Dot Product)을 두 벡터의 크기로 나눈 값으로 계산
  - ✓ 값 범위: -1에서 1 사이
    - 1에 가까울수록 두 벡터가 유사한 방향을 가리키며, 0에 가까울수록 무관함을 의미
    - -1은 두 벡터가 완전히 반대 방향을 의미

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{||A|| \times ||B||}$$

코사인 유사도 (Cosine Similarity)

- **특징**
  - ✓ 평점의 크기보다는 평점의 상대적인 비율 또는 변화의 패턴에 중점
  - ✓ 두 사용자가 아이템에 대해 동일한 크기의 점수를 주지 않더라도 각 아이템에 대한 점수의 높고 낮음이 일관성을 보인다면 코사인 유사도가 높게 계산
    - User1: 영화 1 〉 영화 3 〉 영화 4 〉 영화 2 (높은 점수부터 나열)
    - User2: 영화 1 〉 영화 4 〉 영화 3 〉 영화 2 (높은 점수부터 나열)
    - 두 사용자는 점수 크기가 다르지만 영화 1이 영화 2보다 좋다는 평가 패턴이 유사
    - 사용자가 주는 점수의 크기보다는 각 영화에 대해 더 좋게 평가한 순서가 비슷하면 패턴이 유사하다고 판단

사용자	영화 1	영화 2	영화 3	영화 4
User1	5	3	4	4
User2	3	1	2	3

피어슨 상관계수 (Pearson Correlation Coefficient)

- **정의**
  - ✓ 두 변수 간의 선형 관계를 측정하는 방법
  - ✓ 평균을 기준으로 각 사용자 또는 아이템의 평가 경향성(변화량)을 비교
  - ✓ 주로 사용자 기반 협업 필터링에서 사용
- **계산 방식**
  - ✓ 두 벡터가 평균을 기준으로 얼마나 함께 변동하는지를 측정
  - ✓ 값 범위: -1에서 1 사이
    - 1에 가까울수록 두 변수 간의 완벽한 양의 선형 관계를 의미하고, 0은 무관, -1은 완벽한 음의 선형 관계를 의미

$$r = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum(A_i - \bar{A})^2} \times \sqrt{\sum(B_i - \bar{B})^2}}$$

피어슨 상관계수 (Pearson Correlation Coefficient)

- **특징**
  - ✓ 각 사용자 또는 아이템의 평균적인 평가 경향을 고려하기 때문에 평가 경향이 반영된 유사도 계산이 가능
  - ✓ 단순히 두 사용자가 동일한 아이템에 같은 점수를 부여했는지 뿐만 아니라, 평균보다 높은지 또는 낮은지의 변동 패턴을 측정
    - 영화 1에서는 두 사용자 모두 평균보다 높은 점수를 부여
    - 영화 2에서는 평균보다 낮은 점수를 부여
    - 영화 3과 4에서는 두 사용자 모두 평균과 같은 점수를 부여
    - 각 사용자가 평균적인 평가 경향에 따라 점수를 어떻게 조정했는지를 고려
    - 같은 영화에 같은 점수를 주지 않았더라도 평균보다 높거나 낮은 패턴이 유사하면 높은 상관관계가 있다고 가정

사용자	영화 1	영화 2	영화 3	영화 4
User1	5	3	4	4
User2	3	2	3	3

코사인 유사도 vs. 피어슨 상관계수

- **코사인 유사도 계산**

  ✓ User1과 User2의 코사인 유사도는 0.98로 두 사용자의 평가 패턴은 매우 유사함을 의미

- **피어슨 상관계수 계산**

  ✓ User1과 User2의 피어슨 상관계수는 0.87로 두 사용자의 평가 패턴이 어느 정도 일관된 선형 관계를 가짐을 의미

- **코사인 유사도와 피어슨 상관계수 비교**

  ✓ 코사인 유사도는 0.98로 거의 완벽한 유사성을 보였지만, 피어슨 상관계수는 0.87로 상대적으로 낮은 값을 나타냄

  ✓ 코사인 유사도: 벡터의 방향(패턴)을 중시하기 때문에 두 사용자의 절대적인 평가 값이 다르더라도 패턴이 유사하면 높은 유사도를 제공

  ✓ 피어슨 상관계수: 각 사용자의 평균적인 평가 경향을 반영하므로 사용자가 평가에 일관된 경향을 보이지 않는다면 유사도가 낮게 측정

사용자	영화 1	영화 2	영화 3	영화 4
User1	5	3	4	4
User2	3	1	2	3

선호도 예측

가중평균 (Weighted Average)

- **정의**
  - ✓ 가중평균은 협업 필터링에서 사용자 간 또는 아이템 간의 유사도를 기반으로 가중치를 부여해 평점을 예측하는 방법
  - ✓ 평점에 유사도를 가중치로 곱한 값을 모두 더하고 해당 합을 전체 유사도의 합으로 나누어 평균값을 계산하는 방식
- **공식**
  - ✓ $\hat{r}_{u,i}$: 사용자 u의 아이템 i에 대한 예측 평점
  - ✓ $Sim(u,v)$: 사용자 u와 사용자 v간의 유사도
  - ✓ $r_{v,i}$: 사용자 v가 아이템 i에 남긴 실제 평점

$$\hat{r}_{u,i} = \frac{\sum_{v \in N(u)} \mathrm{Sim}(u,v) \times r_{v,i}}{\sum_{v \in N(u)} |\mathrm{Sim}(u,v)|}$$

- **특징**
  - ✓ 예측 평점이 유사도의 합으로 나누어져 평점이 일정한 범위(예: 1~5) 내에서 계산될 수 있음
  - ✓ 유사도가 높을수록 해당 사용자의 평점이 예측에 더 큰 영향을 미치고 유사도가 낮은 사용자의 평점은 적게 반영됨
  - ✓ 예측 평점은 주로 유사한 사용자들의 행동을 기반으로 계산되므로 유사성이 클수록 예측의 정확도가 높아짐

평균 보정 방식 (Mean-Centered Rating)

- **정의**
  - ✓ 평균 보정은 사용자의 평균 평점 경향을 고려하여 평점 편향을 제거한 후 예측 평점을 계산하는 방식
  - ✓ 사용자마다 평점을 주는 경향이 다르기 때문에 이를 보정하여 예측의 정확도를 높이는 데 사용
- **공식**
  - ✓ $P_{a,i}$: 사용자 a의 아이템 i에 대한 예측 평점
  - ✓ $\bar{r}_a$: 사용자 a의 평균 평점
  - ✓ $r_{u,i}$: 사용자 u가 아이템 i에 남긴 실제 평점
  - ✓ $\bar{r}_u$: 사용자 u의 평균 평점
  - ✓ $w_{a,u}$: 사용자 a와 사용자 u간의 유사도

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \times w_{a,u}}{\sum_{u \in U} |w_{a,u}|}$$

- **특징**
  - ✓ 각 사용자의 평점이 평균보다 얼마나 높은지 혹은 낮은지를 보정하여 개인의 평가 경향성을 제거하고 예측 평점의 편차를 감소
  - ✓ 평균 보정을 통해 예측 평점이 극단적으로 치우치지 않고 평점 범위 내에서 유지
  - ✓ 평균 평점 경향성을 제거하기 때문에 보다 신뢰성 있는 예측이 가능

가중평균과 평균 보정 방식의 비교

구분	가중평균	평균 보정
평가 기준	• 유사도에 따른 가중치를 부여하여 평점 계산	• 평균 평점 경향을 보정한 후 가중평균 방식 적용
평점 편향 제거	• 적용되지 않음	• 사용자 평균을 기반으로 편향 제거
계산 방식	• 유사도에 따라 평점을 가중합한 후, 유사도 합으로 나눔	• 사용자 평균에서 벗어난 정도를 유사도로 가중하여 보정
장점	• 유사도 기반 정규화된 예측 • 단순한 계산	• 편향 제거로 더욱 정확한 예측 제공
단점	• 편향이 제거되지 않아 특정 사용자에게 치우칠 수 있음	• 추가 계산 과정 필요 • 데이터 희소성 문제 시 어려움

성능 평가 지표

예측 성능 평가 지표

- MAE (Mean Absolute Error)
  - ✓ **개념:** 예측된 평점과 실제 평점 간의 절대적인 오차의 평균을 측정하는 지표
  - ✓ 특징: 모든 오차를 동일하게 고려하며 실제와 예측의 평균적인 차이를 측정하는 데 유용
- RMSE (Root Mean Squared Error)
  - ✓ 개념: 예측된 평점과 실제 평점 간의 오차의 제곱 평균을 계산하고 제곱근을 취하는 지표
  - ✓ 특징: 큰 오차에 더 민감하게 반응하며 모델이 큰 오차를 줄이는 데 집중
- MAE와 RMSE의 차이
  - ✓ MAE는 모든 오차를 동일하게 취급하지만, RMSE는 큰 오차에 더 큰 가중치를 부여
  - ✓ 큰 오차를 줄이려면 RMSE를 사용하는 것이 적합

순위 기반 평가 지표

- **Precision (정밀도)**
  - ✓ 개념: 추천된 아이템 중에서 실제로 사용자가 선호한 아이템의 비율을 측정
  - ✓ 특징: 추천된 아이템이 얼마나 정확한지를 평가
- **Recall (재현율)**
  - ✓ 개념: 실제로 사용자가 선호한 아이템 중에서 추천된 아이템의 비율을 측정
  - ✓ 특징: 추천 시스템이 얼마나 많은 사용자의 선호 아이템을 정확하게 찾아냈는지 평가
- **F1-score**
  - ✓ 개념: Precision과 Recall의 조화 평균으로 두 지표 간의 균형을 평가
  - ✓ 특징: Precision과 Recall 중 어느 한쪽에 치우치지 않게 성능을 평가

순위 기반 평가 지표

- MAP (Mean Average Precision)
  - ✓ 개념: 여러 사용자에 대해 추천된 아이템의 Precision 값을 평균내어 성능을 평가하는 지표
  - ✓ 특징: 상위에 추천된 아이템이 사용자의 선호에 얼마나 부합하는지 평가
- NDCG (Normalized Discounted Cumulative Gain)
  - ✓ 개념: 추천된 아이템의 순위를 고려하여 평가하며 상위에 배치된 아이템이 선호될 가능성이 높을수록 높은 점수를 부여
  - ✓ 특징: 추천된 아이템의 순서와 품질을 함께 평가하며 순위에 민감한 지표

협업 필터링의 한계점

콜드 스타트 문제 (Cold Start Problem)

- **정의**
  - ✓ 사용자의 행동 데이터를 기반으로 추천을 제공하는 방식이기 때문에 새로운 사용자 또는 새로운 아이템에 대한 데이터가 충분하지 않을 경우 적절한 추천을 제공하기 어려운 문제가 발생
- **세부 유형**
  - ✓ 신규 사용자 문제: 새로운 사용자가 가입한 경우 해당 사용자의 행동(평가, 클릭, 구매 등)에 대한 데이터가 부족하여 사용자가 좋아할 아이템을 추천하기 어려운 문제가 발생
  - ✓ 신규 아이템 문제: 새로운 아이템이 추가된 경우 해당 아이템에 대한 평가나 상호작용 데이터가 없기 때문에 아이템을 추천하기 어려운 문제가 발생
- **해결 방안**
  - ✓ 하이브리드 추천 시스템을 도입하여 콘텐츠 기반 추천과 협업 필터링을 결합하여 해결할 수 있음
  - ✓ 초기 사용자 정보 수집(예: 선호도 조사, 초기 행동 데이터를 통한 빠른 데이터 확보)을 통해 문제를 완화할 수 있음

데이터 희소성 문제 (Data Sparsity)

- **정의**
  - ✓ 대규모 사용자 및 아이템 집합에서 대부분의 사용자가 일부 아이템에만 평가를 남기기 때문에 평점 행렬이 희소해지는 문제가 발생
  - ✓ 즉, 평점이 비어 있는 셀이 많아 유사도를 계산하거나 추천을 생성하기 어려워지는 문제가 존재
- **문제점**
  - ✓ 데이터가 충분하지 않기 때문에 사용자나 아이템 간의 정확한 유사도 계산이 어려운 문제가 발생
  - ✓ 데이터가 충분하지 않으면 추천 품질이 떨어질 수 있음
- **해결 방안**
  - ✓ 최근접 이웃 기반 방법 대신 모델 기반 방법을 사용하여 희소성을 보완할 수 있음
  - ✓ 행렬 분해 기법(Matrix Factorization)을 사용하여 희소한 평점 행렬을 저차원 공간으로 분해하여 잠재 요인을 추출함으로써 문제를 완화할 수

확장성 문제 (Scalability Problem)

- **정의**
  - ✓ 사용자 수와 아이템 수가 기하급수적으로 늘어날 때 협업 필터링 시스템이 이러한 대규모 데이터에서 효율적으로 유사도를 계산하고 추천을 생성하는 데 어려움을 겪는 문제
- **문제점**
  - ✓ 협업 필터링은 사용자 간 또는 아이템 간의 유사도 계산이 핵심이며, 사용자와 아이템이 많아질수록 유사도를 계산하는 데 필요한 연산량이 급격히 증가
  - ✓ 사용자나 아이템 수가 커질수록 실시간 추천을 제공하는 데 시스템 성능이 저하될 수 있음
- **해결 방안**
  - ✓ 근사화 기법을 사용하여 유사도 계산 시 일부 데이터만을 샘플링하거나 차원 축소 기법을 도입하여 연산 복잡도를 줄일 수 있음
  - ✓ 분산 처리 기술(예: Apache Spark)을 활용하여 대규모 데이터를 병렬로 처리함으로써 확장성 문제를 해결할 수 있음

군집화 문제 (필터 버블)

- **정의**
  - ✓ 협업 필터링은 사용자들의 과거 행동을 기반으로 추천을 제공하기 때문에 사용자가 기존에 선호하던 콘텐츠나 아이템과 유사한 것만 추천받을 가능성이 높음
  - ✓ 이는 사용자가 다양한 콘텐츠를 탐색하지 못하게 하고 특정 콘텐츠나 취향에 고착화되는 문제를 초래할 수 있음
- **문제점**
  - ✓ 협업 필터링은 사용자의 과거 취향을 강화하는 경향이 있으므로 새로운 콘텐츠를 발견할 기회를 제한할 수 있으며 이는 사용자가 다양한 취향을 탐색하지 못하게 함
  - ✓ 사용자에게 이미 선호한 콘텐츠만 계속 추천하게 되어 해당 사용자의 취향이 더 고착화되는 경향이 생길 수 있음
- **해결 방안**
  - ✓ 하이브리드 추천 시스템을 도입하여 협업 필터링과 콘텐츠 기반 필터링을 결합하면 다양한 콘텐츠를 추천할 수 있음
  - ✓ 탐색과 추천의 균형을 유지하는 탐색적 추천 알고리즘(예: 탐험-활용 균형 알고리즘)을 적용하여 다양한 콘텐츠를 추천할 수 있음

잠재적 편향성 문제 (Bias Problem)

- **정의**
  - ✓ 협업 필터링은 기존에 많이 평가된 아이템을 우선적으로 추천하는 경향이 있음
  - ✓ 이로 인해 일부 인기 있는 아이템만 추천되기 쉽고 평가가 적은 아이템은 추천될 기회를 거의 얻지 못하는 문제가 발생
- **문제점**
  - ✓ 협업 필터링은 데이터에 의존하므로 평가된 데이터가 많은 아이템이 추천될 확률이 매우 높아짐
  - ✓ 상대적으로 적게 평가된 아이템은 거의 추천되지 않기 때문에 다양한 아이템을 노출시키는 데 어려움이 존재
- **해결 방안**
  - ✓ 인기 아이템에 대한 가중치 조정이나 평가되지 않은 아이템에 대한 추천을 강화하는 탐색적 추천 알고리즘을 도입할 수 있음
  - ✓ 다양성을 중시하는 추천 시스템 알고리즘(예: 다양성 증진 메커니즘)을 적용하여 더 많은 아이템이 추천될 수 있도록 조정할 수 있음

감사합니다