추천 시스템

목차

1-1 추천 시스템이란?

- 정보 필터링을 기술의 일종으로 특정 사용자가 관심을 가질 만한 정보를 추천

A. 관심을 가질 만한 정보의 예로 영화, 음악, 책, 뉴스, 이미지, 웹 페이지 등

B. 예를 들면 사용자가 특정 곡을 좋아할 경우, 이와 비슷한 콘텐츠를 가진 다른 곡을 추천.

1-1 추천 시스템이란?

- 영화, 음악, 뉴스, 책, 연구주제, 탐색 질의, 상품 등 검색에 적용 가능

- 아마존, 이베이, 유튜브, 전자 상거래 업체 등 다양하게 활용하고 있음.

1-1 추천 시스템이란?

▶ 추천 시스템에 사용되는 데이터 예

- 사용자가 어떤 상품을 구매했을까?
- 사용자가 어떤 상품을 둘러보거나 장바구니에 넣었을까?
- 사용자가 평가한 영화 평점은? 제품 평가는?
- 사용자가 좋아요 한 상품은 무엇인가?
- 사용자가 어떤 상품을 클릭했는가?

1-2 추천 시스템의 유형

■ 추천 시스템의 예

▶ 카카오 웹툰에서의 콘텐츠 기반 필터링

사용자가 카카오 웹툰에서 작품 리스트를 클릭 후, 진입하게 되면 보게 되는 추천 영역

(상세 알고리즘 참조) https://tech.kakao.com/2021/12/27/content-based-filtering-in-kakao/

▶ 콘텐츠 기반 필터링 - Contents-Based Filtering

1-2 추천 시스템의 유형

■ 추천 시스템의 분류

▶ 연관성 규칙 분석 - 장바구니 알고리즘, Association Rules

▶ 협업 필터링 추천 시스템 - Collaborative Filtering

다른 말로 이웃 기반 방식 추천 시스템이라 한다. 또한 유사도 기반 추천 시스템이라고도 한다.

▶ 콘텐츠 기반 필터링 - Contents-Based Filtering

▶ 하이브리드 추천 시스템 - Hybird

좀더 견고한 시스템을 구축하기 위해 다양한 추천 시스템을 결합한다.

다양한 추천 시스템을 결합시키므로 하나의 시스템이 지닌 단점을 다른 시스템의 장점으로 대체 시킨다.

1-2 추천 시스템의 유형

■ 연관성 규칙 분석

‘고객의 장바구니에 함께 담긴 상품들은 서로 연관이 있을 것이다 ' 라고 가정한 후, 이를 토대로 특정 상품을 선택한 고객에게 연관 있는 다른 상품을 추천하는 방법.

1-2 추천 시스템의 유형

■ 추천 시스템의 종류

- 콘텐츠 기반 필터링(Content-based Filtering, CB)
- 협업 필터링(Collaborative Filtering, CF)
  - A. 최근접 이웃(Nearest Neighbor) 협업 필터링(Collaborative filtering)
  - B. 잠재 요인(Latent Factor) 협업 필터링

* 추천 시스템 초창기에는 콘텐츠 기반 필터링이나 최근접 이웃 협업 필터링이 주로 사용되었으나, 넷플릭스 추천 시스템 경연 대회에서 행렬 분해(Matrix Factorization) 기법을 이용한 잠재 요인 협업 필터링 방식이 우승하여 많은 온라인 스토어에서 이를 적용하고 있음.

- 하이브리드 추천 시스템

1-2 추천 시스템의 유형

■ 협업 필터링과 컨텐츠 기반 추천 시스템의 차이

(A) 콘텐츠 기반 필터링은 유저-아이템의 상호작용 데이터가 없어도, 아이템 자체의 정보인 이름, 카테고리, 상세 설명, 이미지 등을 활용하여 유사성을 판단하고 비슷한 아이템 추천이 가능하다.

1-3 협업 필터링(Collaborative Filtering, CF)

■ 협업 필터링(CF)

(A) 추천 모델에서 가장 많이 사용되는 기술이다.

(B) 유저-아이템 간 상호 작용 데이터를 활용하는 방법론

- “이 영화를 좋아했던 사람들은 또 어떤 영화를 좋아할까?”

(C) 핵심 가정은 나와 비슷한 취향을 가진 유저들은 임의의 아이템에 대해 비슷한 선호도를 가질 것이라는 점이다.

1-3 협업 필터링(Collaborative Filtering, CF)

■ 협업 필터링(CF)의 두 가지 접근 방법

(A) 메모리 기반 접근 방식

유저 간/아이템 간 유사도를 메모리에 저장한 후, 특정 유저에 대해 추천이 필요할 때 해당 유저와 유사한 k명의 유저가 소비한 아이템을 추천하거나, 혹은 특정 아이템에 대한 Rating 예측이 필요할 때 해당 아이템과 유사한 k개의 아이템의 Rating을 기반으로 추정이 가능.

종류 : 사용자 기반 협업 필터링(User-based Filtering), 아이템 기반 협업 필터링(Item-based Filtering)

(B) 모델 기반 접근 방식

(1) Latent Factor방식

(2) Classification/Regression 방식 및 딥러닝을 사용한 접근

1-3 협업 필터링(Collaborative Filtering, CF)

■ 모델 기반 접근 방식 – Latent Factor 방식

(1) Latent Factor방식

(2) Classification/Regression 방식 및 딥러닝을 사용한 접근

1-3 최근접 이웃 협업 필터링

▶ 최근접 이웃 협업 필터링

- 자기와 취향이 같은 친구들이 무엇을 샀는지 물어보기

▶ 종류

A. 사용자 기반 협업 필터링

B. 아이템 기반 협업 필터링

1-3 최근접 이웃 협업 필터링

▶ 사용자 기반(User-User) 협업 필터링

- 사용자-사용자 간의 유사도를 측정

▶ 아이템 기반(Item-Item) 협업 필터링

- 아이템-아이템 간 유사도를 측정

1-3 최근접 이웃 협업 필터링

▶ 사용자 기반 협업 필터링

A. 특정 사용자와 타 사용자 간의 유사도(Similarity)를 측정한 뒤 가장 유사도가 높은 TON-N 사용자를 추출해 그들이 선호하는 아이템을 추천하는 것이다.

1-3 최근접 이웃 협업 필터링

▶ 사용자 기반 협업 필터링

	Item A	Item B	Item C	Item D	Item E
사용자 A	5	5	3	2	5
사용자 B	4	5	4	2	
사용자 C	2	2	3	2	

- A는 아이템 a, b, c, d에 5,5,3,2, e에 5점을 주었다.

- B는 아이템 a, b, c, d에 4,5,4,2 라는 평점을 주었다.

- A, B는 사용자 간 유사도가 높기 때문에 A가 좋아한 아이템 E를 B에게 추천한다.

1-3 최근접 이웃 협업 필터링

▶ 아이템 기반 협업 필터링

	사용자 A	사용자 B	사용자 C	사용자 D	사용자 E
죽은 시인의 사회	5	5	4	4	5
굿 윌 헌팅	4	5	4	3	
어벤져스	3	3	2	3	

- 사용자 A, B, C, D, E는 죽은 시인의 사회의 영화에 5,5,4,4,5를 평점을 주었습니다.

- 사용자 A, B, C, D는 굿 윌 헌팅의 영화에 4,5,4,3을 주었습니다.

- 죽인 시인의 사회를 좋아한 사용자 E에게 굿 윌 헌팅을 추천해 줍니다.

1-3 최근접 이웃 협업 필터링

▶ 아이템 기반 협업 필터링

$$\hat{R}_{u,i} = \sum_N (S_{i,N} * R_{u,N}) / \sum_N (|S_{i,N}|)$$

$\hat{R}_{u,i}$: 사용자 u, 아이템 i의 개인화된 예측 평점 값

$S_{i,N}$: 아이템 i와 가장 유사도가 높은 Top-N개 아이템의 유사도 벡터

$R_{u,N}$: 사용자 u의 아이템 i와 가장 유사도가 높은 Top-N개 아이템에 대한 실제 평점 벡터

```
def predict_rating(ratings_arr, item_sim_arr):
    val = np.array(  [np.abs(item_sim_arr).sum(axis=1)]  )  # 유사도의 열의 합
    pred = ratings_arr.dot(item_sim_arr) / val
    return pred
```

1-3 최근접 이웃 협업 필터링

▶ 아이템 기반 협업 필터링 – 유사도 계산하기

- Cosine-Based Similarity (코사인 기반 유사성)
- Correlation-Based Similarity(상관 계수 기반 유사성)
- Adjusted Cosine Similarity(수정된 코사인 기반 유사성)
- 1-Jaccard distance(1-자카드 거리)

1-3 최근접 이웃 협업 필터링

▶ 장단점

- [장점] 구현하기 쉽다.
- [장점] 추천 생성 시에 제품의 콘텐츠 정보 또는 사용자 프로필 정보는 필요하지 않는다.
- [장점] 사용자에게 깜짝 놀랄 만한 새로운 아이템을 추천한다.
- [단점] 유사도 계산을 위해 모든 사용자, 제품, 그리고 등급 정보가 메모리에 로드된다.
- [단점] 데이터가 거의 없는 경우 성능이 저하된다.
- [단점] 사용자 또는 제품에 대한 콘텐츠 정보가 없기 때문에 등급 정보만으로는 정확한 추천을 생성할 수 없다.

1-4 콘텐츠 기반 필터링 추천 시스템

■ 콘텐츠 기반 필터링

(A) 사용자가 특정한 아이템을 많이 선호하는 경우,

그 아이템과 비슷한 콘텐츠를 가진 다른 아이템을 추천하는 방식

⇒ 사용자가 특정 영화에 높은 평점을 주었다면 그 영화의 장르, 출연 배우, 감독, 영화 키워드 등의 콘텐츠와 유사한 다른 영화를 추천해 주는 방식

(B) 아이템이 유사한지 확인하기 위해 아이템의 비슷한 정도(유사도, similarity)를 수치로 계산할 수 있어야 한다.

⇒ 유사도 계산을 위해 일반적으로 아이템을 벡터 형태로 표현. 이들 간의 벡터 간의 유사도 계산 방법을 많이 활용.

⇒ 아이템을 잘 표현할 수 있는 벡터를 만들기 위해 원 핫 인코딩, 임베딩의 사용 가능.

1-4 콘텐츠 기반 추천 시스템

▶ 장점

- 협업 필터링처럼 사용자 커뮤니티를 이용하기보다 사용자 선호도만을 이용해서 추천 생성.

- 등급 정보만을 처리하는 것이 아니고, 제품의 콘텐츠를 처리하기 때문에 협업 방식 대비 정확도가 높다.

- 이러한 접근법은 추천 사항을 처리 또는 생성하기 위해 추천 모델이 모든 데이터를 로드할 필요가 없으므로 실시간으로 적용 가능.

1-4 콘텐츠 기반 추천 시스템

▶ 단점

- 시스템이 더 개인화됨에 따라 더 많은 사용자 정보가 시스템으로 유입되면 오직 사용자 선호도에만 집중된 추천 사항을 생성

- 사용자 선호도와 관련 없는 새로운 제품은 사용자에게 소개하지 않는다.

- 사용자는 자신의 주변에서 일어나는 일이나 최신 트렌드를 파악할 수 없게 된다.

1-5 유사도 측정 방법

▶ 유사도 측정 방법

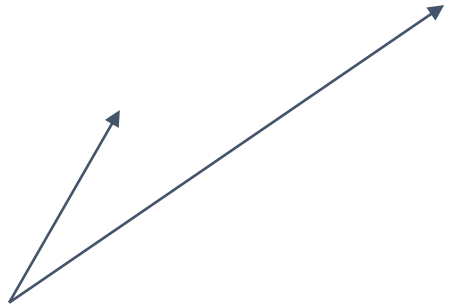
A. 거리 측정 방법은 다양

- Cosine Similarity
- Pearson Similarity
- Euclidean Distance

1-5 유사도 측정 방법

▶ 코사인 유사도의 상대적 장점

좌표상에 거리로만 우리가 관계가 있다 없다 표현할 경우, 같은 방향이지만, 크기가 다를 경우, 이는 관계가 없다고 판단한다면 오류를 범할 수 있다.



- 우리는 이 경우, 데이터가 이루는 사잇각 θ로 유사도를 측정할 수 있다. θ가 작으면 데이터의 유사도가 높고, θ가 크면 데이터의 유사도가 낮다고 판단할 수 있다. 사잇각이 0이면 유사도는 1
- 그러나 사잇각은 벡터의 내적(inner product)로부터 정의되므로 θ를 직접 계산하기 보다 벡터의 내적을 이용하여 θ의 코사인 값으로 유사도를 측정한다. 이를 코사인 유사도(cosine similarity)라 한다.

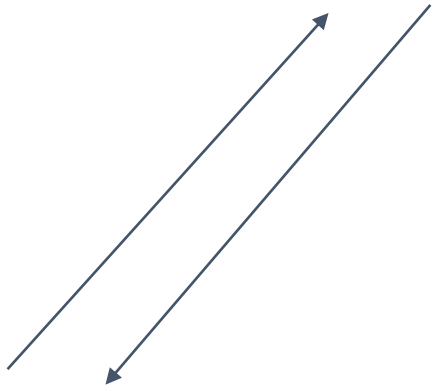
1-5 유사도 측정 방법

▶ 코사인 유사도

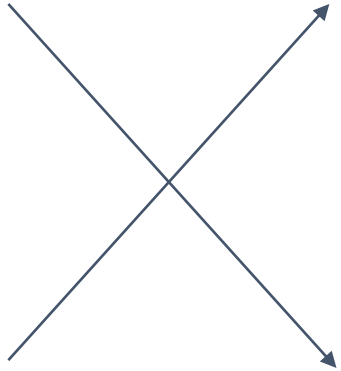
* 코사인 유사도는 두 벡터 간의 각도를 이용하여 구할 수 있는 두 벡터의 유사도를 의미.

* 코사인 유사도는 -1 ~ 1 이하의 값

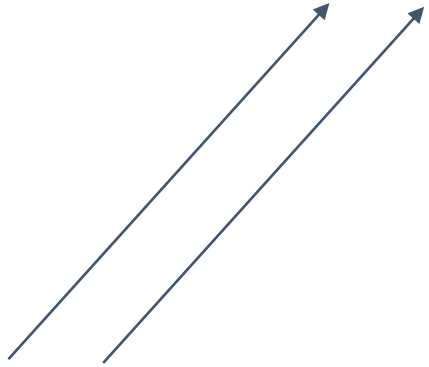
* 코사인 유사도는 두 벡터 간의 사잇각을 코사인 씌워준 값을 통해 구할 수 있다.



▶ 코사인 유사도 : -1



▶ 코사인 유사도 : 0



▶ 코사인 유사도 : 1

1-5 유사도 측정 방법

▶ 코사인 유사도

$$A \cdot B = |A| * |B| * \cos\theta$$

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

1-6 잠재 요인(Latent Factor) 협업 필터링

▶ 잠재 요인 협업 필터링

A. 행렬 분해(matrix Factorization)을 기반하여 사용.

- 다차원 매트릭스를 저차원 매트릭스로 분해하는 기법 SVD, NMF등이 있다.

B. 사용자-아이템 평점 매트릭스 속에 숨어 있는 잠재 요인을 추출해 추천 예측을 할 수 있도록 하는 기법.

C. 사용자-아이템 평점 행렬 데이터만을 이용해 말 그대로 ‘ 잠재 요인’을 끄집어 내는 것을 의미

1-7 Classification/Regression(분류/회귀) 방식

- 피처 x가 주어졌을 때, 라벨 y를 에측한다.
- 콘텐츠 관련 정보를 x로 두고, 이에 따라 y를 추천이 가능.

REFERENCE

- 카카오 추천팀 Github

  https://github.com/kakao/recoteam