메모리 기반 협업 필터링을 활용한 웹 기사 추천

2024.07.05

김태형, 이준걸

- 1. 문제 정의 및 데이터 설명
- 2. 데이터 전처리
- 3. 방법론
- 4. 실험 결과 및 결론

1. 문제 정의 및 데이터 설명

웹 기사 추천

- 웹 기사 조회 로그 데이터를 기반으로 사용자에게 맞춤형 기사를 추천하는 AI 알고리즘을 개발함.
- 데이터 설명은 다음과 같음.
- view_log.csv train 데이터
 - 유저가 기사를 조회한 로그 데이터, 학습 데이터이며 해당 데이터에 존재하는 유저만 추천의 대상이 됨.
 - userID : 유저 고유 ID, articleID : 기사 고유 ID
 - userRegion : 유저가 속한 지역, userCountry : 유저가 속한 국가
- article_info.csv meta데이터
 - 기사에 대한 정보
 - articleID: 기사 고유 ID, Title: 기사의 제목, Content: 기사의 본문
 - Format : 기사의 형식, Language : 기사가 작성된 언어
 - userID: 기사를 작성한 유저 고유 ID, view_log에 포함되지 않은 유저가 존재할 수 있으며, 해당 유저는 추천의 대상이 되지 않음.
 - userCountry: 기사를 작성한 유저가 속한 국가, userRegion: 기사를 작성한 유저가 속한 지역
- sample_submission.csv 제출 양식
 - 한 유저에게 5개의 기사를 추천하게 되며, 유저가 기존에 조회한 기사 추천 가능

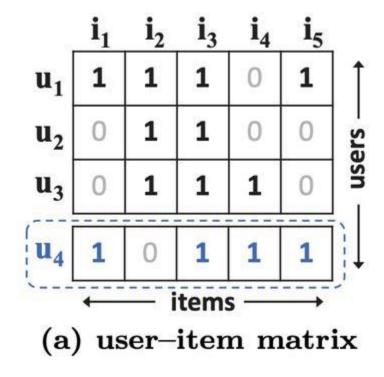
2. 데이터 전처리

탐색적 데이터 분석

- view_log_train에서 관측된 user는 1415명, article은 2879임.
- article_info에서 관측된 전체 article은 3008개이나 user 기록이 존재하는 2879개의 article만 추천하고자 함.
- view_log_train에서 관측되지 않고 article_info에서만 관측된 추가 user는 9명이나 추가 user는 추천의 대상이 되지 않음.
- 해당 문제는 Cold Start Problem, Long Tail, Efficiency를 고려하지 않아도 된다고 판단됨.
- 때문에, Memory-based Collaborative Filtering을 사용함.
- User-based Collaborative Filtering과 Item-based Collaborative Filtering을 가중 평균하여 모델링함.
- 유저가 기존에 조회한 기사를 추천할 수 있기 때문에, 명시적으로 유저가 많이 조회한 기사를 추천하는 규칙 기반의 알고리즘을 설계함.
 - ① view_log_train을 기반으로 각 사용자가 조회한 기사 리스트를 추출함.
 - ② 사용자가 조회한 기사가 5개 미만일 때, 추천 점수가 높은 순서대로 남은 기사를 추천함.

3. 방법론

Memory-based Collaborative Filtering



- 협업 필터링은 사용자의 구매 패턴이나 평점을 가지고 다른 사람들의 구매 패턴, 평점을 통해서 추천하는 방법임.
- 추가적인 사용자의 개인정보나 아이템의 정보가 없이도 추천할 수 있는게 큰 장점임.
- 본 경진대회에서는 User 기반 또는 Item 기반의 Cosine Similarity를 추천 점수로 활용함.

4.실험 결과 및 결론

Experiments & Results

- 최종 제출 모델은 User-Item-based collaborative Filtering 모델 (alpha=0.3)임.
- Cross-Validation을 사용하여 하이퍼 파라미터 alpha(0.3)를 탐색함.
- User-based Collaborative Filtering 모델의 public Recall@5는 0.34965이며 Private Recall@5는 0.34142임.
- Item-based Collaborative Filtering 모델의 public Recall@5는 0.34066이며 Private Recall@5는 0.32980임.
- User-Item-based collaborative Filtering 모델 (alpha=0.5)의 Public Recall@5는 0.34846이며 Private Recall@5는 0.34임.
- User-Item-based collaborative Filtering 모델 (alpha=0.3)의 Public Recall@5는 0.35280이며 Private Recall@5는 0.34142임.
- 추가적으로 Neural Collaborative Filtering, LightGCN 등의 딥러닝 기반의 추천시스템 모델을 구현했으나 성능을 향상시키지 못함.
- Sentence Transformer를 활용하여 Item Embedding을 진행하였으나 성능을 향상시키지 못함.
- 명시적으로 유저가 많이 조회한 기사를 추천하는 규칙 기반의 알고리즘이 성능 향상에 크게 도움됨.

Thank you ©

taehyeong93@korea.ac.kr