Multifaceted Collaborative Filtering Model

Domain	RecSys
tag	Integration of implicit & explicit Integration of neighborhood & latent model
Conference / Journal	KDD
Publish year	2008
정리 날짜	@2024년 1월 12일
AI summary	This document discusses the integration of neighborhood models and latent factor models in collaborative filtering for recommendation systems. It explores the motivation behind using both explicit and implicit feedback, and introduces an integrated model that combines the strengths of both approaches. The methodology includes baseline estimates, neighborhood models using Pearson correlation coefficients, and latent factor models using SVD. The conclusion shows that the integrated model leads to a decrease in RMSE and demonstrates improved performance in top-k evaluation.
AI key info	RecSys, Content-based filtering, Collaborative filtering, Neighborhood Model, Latent factor model, Feedback, RMSE, Top-K recommender, Baseline estimate, Pearson correlation coefficient, Shrunk correlation coefficient, Interpolation weight,

	SVD, Alternative Least Square Method, SGD, Asymmetric-SVD, SVD++

Summary

Latent model과 neighborhood model을 통합해 Collaborative Filtering하는 방법 제시

- Implicit feedback, explicit feedback 모두 활용

Background

RecSys

- Content-based filltering
    - item, user profile을 만드는 것이 핵심
        - item의 feature을 모두 설정해줘야 함
            - 반영되지 않는 feature 있을 수 있음
        - 관계에 대해 학습이 되지 않을수도
- Collaborative filltering
    - 개별 사용자의 과거 행동 정보 뿐 아니라, 다른 사용자들과의 관련성도 사용하여 filtering 함
    - Directly rely on user behavior
        - Explicit profile 만들 필요 없음
            - extensive data collection 필요 X
    - Memory-based CF
        - user based / item based: 비슷한 user/item 기반으로 추천
        - e.g. Neighborhood Model
            - Item간, user간 relationship 계산
            - localized relationship에 특화
                - totality of weak signal을 반영하기 어려움
            - Item oriented가 최근 더 핫함

        - better scalability, improved accuracy, explaining reasons of prediction
    - e.g. kNN: 가까운 k개 user와 비슷한 content 추천
- Model-based CF
  - e.g. Latent factor model
    - User와 item을 잠재 변수인 vector들로 나타내는 것
      - 모든 feature를 모두 일일이 표현할 필요 없음
        - Uninterpretable dimension이 잠재 변수로 모두 표현되지 않을까
      - User-Item interaction의 전체 행렬을 고려하기 때문에 간접적으로 타 user 활용
    - Estimate overall structure에 강함
      - strong association among a small set of closely related 약함
    - e.g. SVD
- Feedback
  - explicit feedback
  - implicit feedback

Motivation

- Neighborhood model과 latent factor model 모두 사용해서 학습
  - 두 모델의 장점을 하나의 모델에 녹여보자
- Explicit, implicit feedback 모두 사용해보자
- RMSE를 높이는 것이 과연 효과가 있는가
  - Top-K recommender 도입

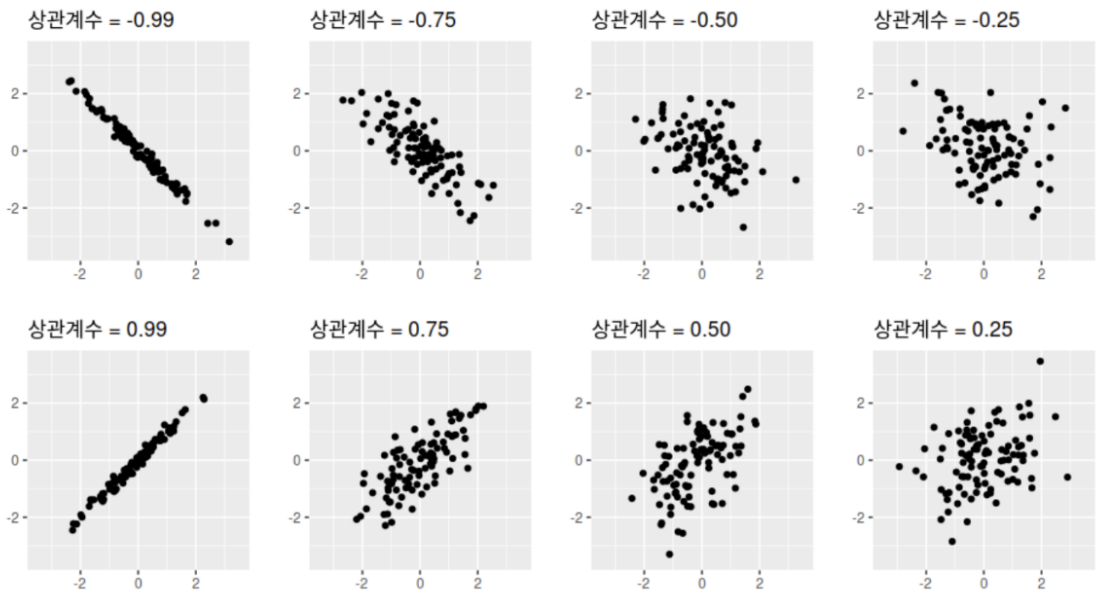
Methodology

Preliminary

- Baseline estimate
  - 각 유저, item들의 tendency 반영하기 위해 baseline estimate 도입

$$b_{ui} = \mu + b_u + b_i$$

- Neighborhood model
  - Pearson correlation coefficient : item-oriented approach의 흔한 similarity measure
    - Tendency of users to rate items similarly



  - Shrunk correlation coefficient $s_{ij} \overset{\text{def}}{=} \frac{n_{ij}}{n_{ij}+\lambda_2}\rho_{ij}$
    - item을 평가하는 사람이 있고 안하는 사람이 있으니, 평가수가 많을수록 relable로 판단
      - 평가수가 많아질수록 pearson coefficient에 수렴함
    - λ_2는 일반적으로 100으로 설정
  - rating r_{ui} 예측

        - k개 neighbor의 상관계수 고려하여 rating 계산해 baseline에 더한 값

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in S^k(i;u)} s_{ij}(r_{uj} - b_{uj})}{\sum_{j \in S^k(i;u)} s_{ij}}$$

        - 문제점
            - formal model로 정당화되지 않음
            - 전체 neighbor들의 interaction 고려하지 않고, 두개 item을 isolate함
            - neighbor info가 없을 때에도 neighbor에만 의지함
    - 대안: Interpolation weight 계산

$$\hat{r}_{ui} = b_{ui} + \sum_{j \in S^k(i;u)} \theta_{ij}^u (r_{uj} - b_{uj})$$

- Latent factor model
    - SVD를 주로 사용
        - user-factor vector와 item-factor vector로 분해

$$\hat{r}_{ui} = b_{ui} + p_u^T q_i.$$

        - CF는 missing value가 많아 그대로 적용하면 문제
        - overfitting에 취약함
            - 예전에는 imputation으로 데이터 양 늘렸지만, 데이터 크기, 왜곡 문제 등 발생
            - Regularization으로 해결

$$\min_{p_*,q_*,b_*} \sum_{(u,i)\in\mathcal{K}} (r_{ui}-\mu-b_u-b_i-p_u^T q_i)^2+\lambda_3(\|p_u\|^2+\|q_i\|^2+b_u^2+b_i^2)$$

    - 단순하게 gradient descent 적용 가능

Interpolation weight model

- explicit 뿐 아니라 implicit도 모델 식에 반영
- normalization도 도입

학습 방법

- Alternative Least Square Method
- SGD

Asymmetric-SVD

SVD++

- SVD에서 implicit도 사용

Conclusion

Integrated model

- neighborhood model, latent fator model의 수식을 단순합한 것
- RMSE가 0.07감소하는 효과
    - 큰 감소 아니지만, top-k evaluation을 통해 이 수치가 의미 있는 성능 개선을 보임을 증명
        - 평점 높은 item 실제 높게 예측률