개인화 추천시스템의 사용자 선호도를 반영한 데이터 통합 방법

Knowledge-based Data Discovery
INHA KDD Lab.

2021.11.25

여민영
인하대학교
dufmasky@gmail.com

목 차

서론

- **기존 추천시스템 한계점**
  - Explicit data 단독 사용시 데이터 수집에 어려움이 있을 수 있고, Cold start problem과 점수 편향의 문제 발생 가능
  - Implicit data 단독 사용시 잡음이 많으며, 데이터 활용을 위한 적절한 평가척도 필요
  - 단독으로 사용할 시 발생할 수 있는 문제점 때문에 요즘은 **데이터셋을 통합 (Explicit data + Implicit data) 하는 추세**
  - 통합 데이터에 대한 **명확한 평가척도 부재**

- **본 연구의 목적 및 필요성**
  - Explicit data와 Implicit data를 **혼합한 정제된 데이터셋** 제작
  - 해당 데이터셋에 대한 **적절한 평가척도** 고안
  - 모델 기반 협업 필터링 알고리즘을 통한 원본 데이터셋(리뷰데이터)과 **효율성 비교 및 검증**

2. 연구 설계

실험 과정

- 실험 과정
  - 여성건강 생활케어 플랫폼인 '먼슬리싱' 기업의 데이터를 제공받아 그림 1에서 도식화된 과정에 따라 실험진행
  - 데이터셋은 어플에서 판매되는 제품들에 대한 1,960개의 리뷰데이터와 장바구니, 구매여부가 저장된 43,758개의 데이터 포함
  - **그림1의 Data handling 과정에서 본 실험에 필요한 데이터를 추출하여 New rating 데이터셋 제작**
  - MF 알고리즘인 **SVD, SVD++, NMF**와 **ALS**를 사용하여 성능비교

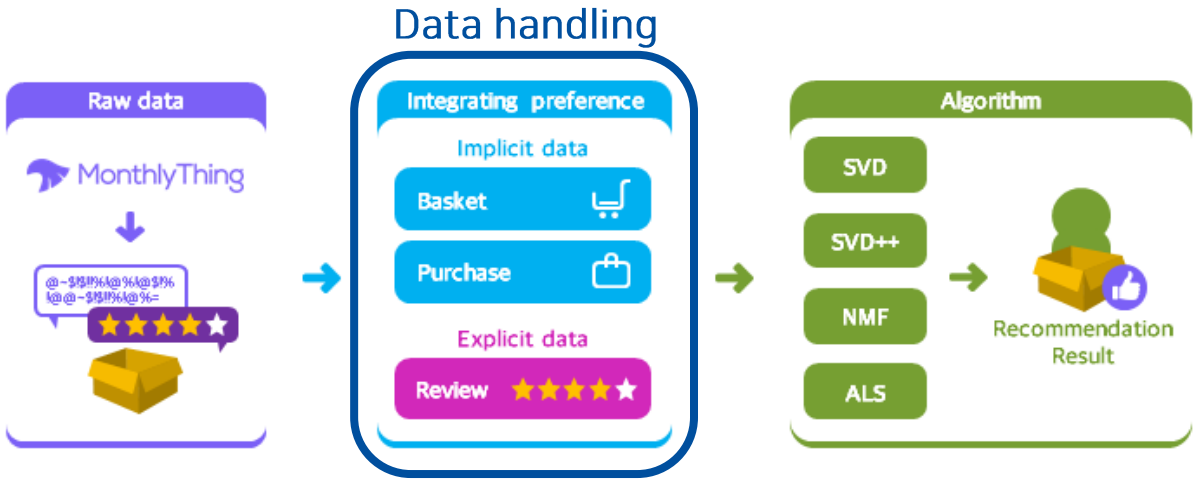

그림 1. 실험 과정

2. 연구 설계

통합데이터셋 점수 기준

- 새로 제작한 통합데이터셋 점수의 기준
  - Original rating으로 사용된 리뷰데이터는 0.1점 단위로 표현되는 0~5점의 값
  - **Threshold값**을 지정하여, 단순히 Rating을 합하여 수식을 결정할 경우 Basket(장바구니), Purchase(구매), Review(평점) 사이에 발생할 수 있는 **우선순위 제거**, 그리고 **제품의 선호도를 반영**
  - New rating에서 사용된 리뷰데이터는 Original rating에서 쓰인 리뷰데이터에서 Threshold값을 빼 준 값
  - Basket과 Purchase는 각각 장바구니와 구매 여부를 나타낸 **Binary한 값**

Category	Original rating	New rating		
Rating point	Review	Basket	Purchase	Review
	1~5	No : 0 / Yes : 1	No: 0 / Yes:1	1~5 -Threshold
Count	1,657	45,716	19,326	1,657
Total	1,657	66,699		

표 1. 통합 데이터셋의 점수 기준

통합데이터셋 구체화 및 정규화

- **Threshold값 지정 후 Rating 예시와 근거**
  - 장바구니O, 구매O, 리뷰 3점 : 1+1+(3-3) = 2점 – 선호 o
  - 장바구니O, 구매O, 리뷰 1점 : 1+1+(1-3) = 0점 – 선호 x
  - 장바구니O, 구매x, 리뷰 x : 1+0+0 = 1점 – 선호 o
  - Explicit 데이터셋을 그대로 반영하여 데이터의 손실이 없음
  - 통합데이터(Explicit+Implicit)의 Rating을 정하는 과정에서 두 데이터 간 명확한 비교를 위한 새로운 **Rating** 방식을 제안
  - Threshold는 이후 F1 Measure을 구할 때 재활용

통합 데이터 / Explicit 데이터

$$r_{ui} = \left(b_{ui}^{basket} + b_{ui}^{purchase}\right) + \left(r'_{ui} - T\right)$$

* $b_{ui,\,basket}$: 장바구니 여부 0/1
* $b_{ui,\,purchase}$: 구매 여부 0/1
* r'_{ui}: user rating
* r_{ui}: new rating
* T: threshold

수식 1. 통합 데이터셋의 Rating

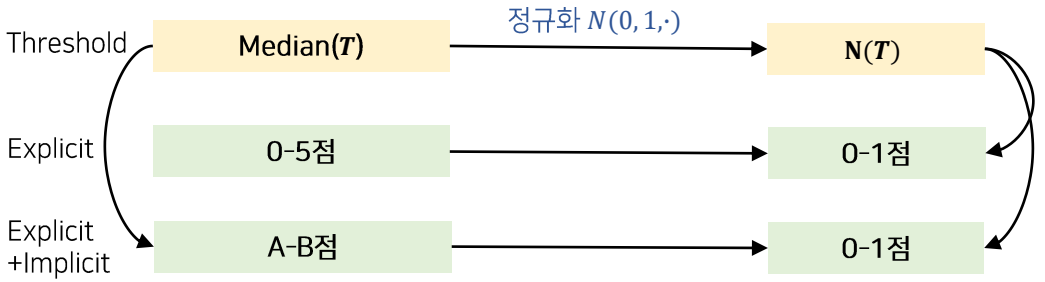

그림 2. 통합 데이터셋 정규화

연구 결과

- 평점 예측 기반 성능평가 결과
  - 그림 3에 의하면, 통합데이터가 모든 알고리즘에서 학습 결과, 수치가 평균적으로 약 33.0% 감소
  - 네 개의 모델 중, SVD++모델이 RMSE 기준 0.1019, MAE기준 0.078로 가장 뛰어남
- Ranking 기반 성능평가 결과
  - 그림 4에 의하면, 통합데이터가 모든 알고리즘에서 학습결과가 개선
  - F1점수에 대해 평균적으로 36.0% 증가하였으며 SVD++모델이 0.68로 가장 높음

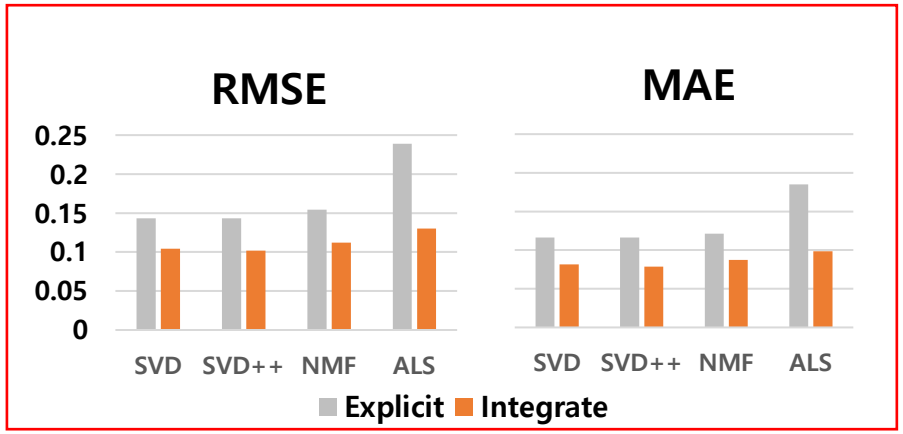

그림 3. MF 기반 예측 알고리즘의 회귀 평가지표

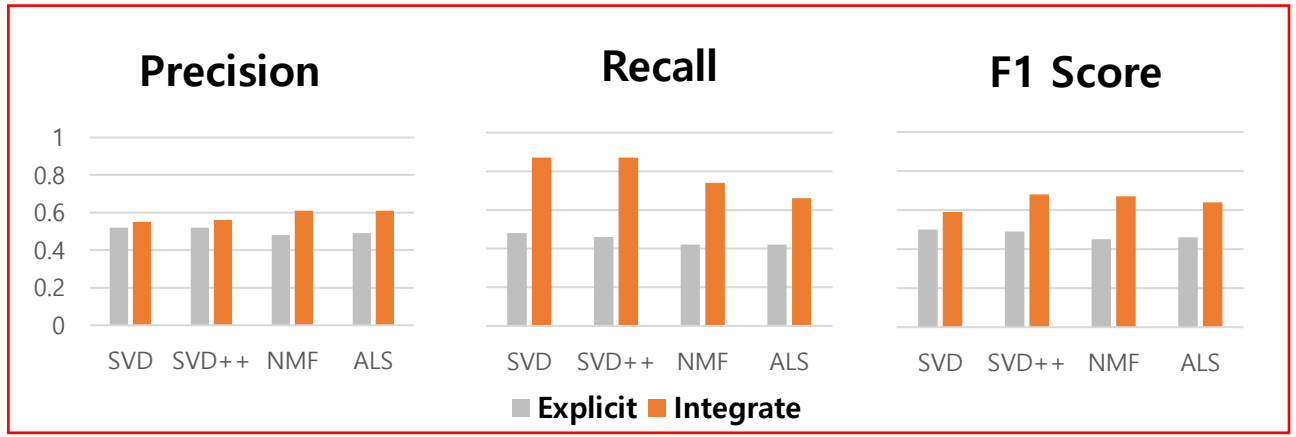

그림 4. MF 기반 추천결과에 따른 성능평가

결론 및 향후 연구

- 결론
  - 본 연구는 명시적 리뷰 데이터와 암묵적 데이터를 결합한 새로운 통합 데이터셋 제작
  - 통합데이터의 Rating을 정하는 과정에서 데이터 간 명확한 비교를 위한 새로운 선호도 수식 제안
  - 평점 예측 기반 성능평가 결과, RMSE와 MAE 기준으로 수치가 평균적으로 33.0%감소
  - Ranking 기반 성능평가 결과, F1 Score 기준으로 수치가 평균적으로 36.0% 증가
  - 다양한 MF 알고리즘에 대한 전반적인 성능 개선

- 향후 연구
  - 다양한 암묵적 데이터를 반영하기 위하여 일반화된 가중치와 새로운 선호도 수식 검증 예정
  - 웹 크롤링을 통해 e-커머스에 있는 리뷰데이터를 확보하여 정확도 상승