아이템 정보 기반 협업 필터링 추천 시스템 연구

양영욱*, 윤유동*, 임희석*
*고려대학교 정보대학 컴퓨터학과
e-mail:yeongwook@blp.korea.ac.kr

A Study on Collaborative Filtering Recommender system based on Item Knowledge

Yeong-Wook Yang*, You-Dong Yun*, Heui-Seok Lim*
*Department of Computer Science and Engineering, Korea University

요 약

Matrix factorization은 사용자의 아이템 선호도를 통해 아이템을 추천해주는 성공적인 기술 중 하나이다. 이 기법은 사용자-아이템의 선호도 행렬을 채우는 것을 목표로 한다. 이 목표를 달성하기 위해 사용자-아이템의 선호도 행렬을 사용자 행렬(user latent factor)와 아이템 행렬(item latent factor)로 분해하고, 각 행렬에 대해 추론하여 완성된 사용자-아이템의 선호도 행렬을 추론한다. 하지만 Matrix factorization은 아이템의 수가 많고, 아이템에 대한 사용자들의 선호도 데이터가 적을 때 성능이 제한된다. 또한 새로운 아이템이 추가되었을 때, 새로운 아이템에 대한 사용자들의 선호도 정보가 없기 때문에 새로운 아이템이 추천되지 않는다는 문제를 가진다. 이를 해결하기 위해 본 논문에서는 아이템에 대한 부가적인 정보인 아이템 간의 유사도 정보와 아이템의 시나리오 정보의 유사도를 모델링하여 기존의 전통적인 Matrix factorization에 추가하는 아이템 정보 기반 추천 시스템을 제안한다.

1. 서론

추천 시스템은 사용자들이 소비한 아이템에 대한 선호도를 통해서 사용자를 모델링한다. (사용자, 아이템, 선호도)의 셋으로 사용자가 가지는 아이템 선호도를 표현하며, 사용자들의 선호도 정보를 통해 사용자가 소비할 아이템을 예측한다. 예를 들어, 사용자 A가 아이템 B에 대해 4점이라는 선호도 점수를 주었다면, 사용자가 평가하지 않은 유사한 아이템에 대해 추천해 주는 것이다.

사용자의 아이템 선호도를 통해서 사용자에게 아이템을 추천해 주는 성공적인 기술 중 하나는 matrix factorization(MF)이다[1]. MF는 사용자의 아이템 선호도를 행렬로 표현하고, 모든 선호도를 추정하는 것을 목적으로 한다. 즉, 행렬의 비어있는 부분을 채우는 것이다. 이 방법은 대게 높은 성능을 보인다[1].

하지만 아이템이 많고, 아이템에 대한 사용자들의 선호도 데이터가 적을 때, 추천에 대한 성능이 제한된다. 이 문제를 data sparsity 문제라고 한다. 또한 새로운 아이템이 추가되었을 때, 새로운 아이템에 대해 사용자들의 선호도 정보가 없기 때문에 새로운 아이템은 추천되지 않는다는 문제점을 가지고 있다. 이 문제는 cold-start 문제라고 한다. 이러한 문제점들을 극복하기 위해 아이템에 대한 부가적인 정보를 추가한 추천시스템 구성을 통해 더 좋은 성능을 달성할 수 있다[2].

본 연구에서는 아이템에 대한 정보로 아이템 간의 유사도와 아이템에 대한 시나리오 정보를 추가하여 추천 모델을 학습하는 방법을 제안한다. 이 방법을 통해 data sparsity 문제와 cold-start문제를 극복하는 모델을 제안한다.

2. 관련연구

Perm Gopalan 외 1명은 아이템에 대한 사용자의 선호도를 예측하기 위해 Poisson 분포를 사용하여 사용자의 행동과 아이템에 대한 텍스트를 모델링하였다[3]. Julian McAuley 외 1명은 아이템에 대한 리뷰 데이터를 통해 아이템이 가지고 있는 주제 정보를 모델링하고, 주제 정보를 전통적인 모델에 추가하여 성능을 향상 시켰다[4]. Chong Wang 외 1명은 아이템이 가지고 있는 단어 정보를 통해 주제 정보를 모델링하고, 아이템의 선택 정보를 통해 아이템에 대한 사용자의 선호도를 모델링하였다[5]. 추천 모델의 성능을 높이기 위해 아이템과 관련된 부가적인 정보를 전통적인 모델에 추가하여 성능을 향상시켰다.

본 연구에서는 아이템에 대한 부가적인 정보로 아이템 간의 유사도와 아이템의 시나리오 정보를 통한 유사도를 모델링하여 기존의 MF에 추가하여 아이템 정보 기반 추

천 모델을 구성하였다.

3. 아이템 정보 기반 추천 모델

아이템 정보 기반 추천 모델은 기존의 MF에 부가적인 아이템 정보를 추가하여 추천 모델을 정규화한다.

3.1. Matrix factorization

Matrix factorization(MF)은 사용자와 아이템에 대한 사용자-아이템 행렬이 주어졌을 때, 사용자 행렬(user latent factor)과 아이템 행렬(item latent factor)로 행렬($U \times I$)을 분해한다. 내적 피드백 모델에 대한 MF는 아래와 같이 표현될 수 있다.

$$L = \sum_{u,i} c_{ui}(y_{ui} - \theta_u \beta_i)^2 + \lambda_\theta \sum_u \| \theta_u \|_2^2 + \lambda_\beta \sum_i \| \beta_i \|_2^2 \quad (1)$$

c는 스케일링 파라메터이고, 보통 $c_{y=1} > c_{y=0}$로 설정한다. y는 관측된 아이템에 대한 선호도 행렬로 $U \times I$의 차원을 갖는다. θ는 user latent factor이고, β는 item latent factor를 의미한다. λ_θ와 λ_β는 정규화 파라메터이다. user latent factor와 item latent factor를 추론함으로 L에 대한 최적의 해를 예측한다.

3.2. 아이템 정보 모델

아이템에 대한 사용자의 선호도를 보충할 수 있는 것은 아이템에 관한 정보이다. 본 논문에서는 아이템 간의 유사도 정보와 아이템의 시나리오 정보를 사용한다.

아이템 간의 유사도는 사용자의 아이템에 대한 평가 정보를 사용하여 아이템 벡터간의 유사도를 측정한다. 본 논문에서는 코사인 유사도 방식을 사용하며, 두 아이템 간의 유사도는 아래와 같이 표현할 수 있다.

$$\cos(p_i, p_j) = \frac{\sum_{k=1}^{m} v_{ki} v_{kj}}{\sqrt{\sum_{k=1}^{m} v_{ki}^2 \sum_{k=1}^{m} v_{kj}^2}} \quad (2)$$

m은 사용자의 수를 의미하며, p_i와 p_j는 두 아이템을 의미하고, v는 벡터를 의미한다.

아이템의 시나리오 정보는 단어의 빈도수를 통하여 아이템 간의 유사도를 측정한다. 유사도를 측정하는 방식은 위의 수식과 동일하며, 사용자의 평가 정보 대신 아이템에서 나타나는 단어의 빈도를 통해 아이템 간의 유사도를 측정한다. 아이템에서 나타나는 단어의 종류와 빈도가 유사할수록 아이템간의 유사도가 높은 결과로 도출된다.

아이템 간의 유사도 행렬과 아이템 시나리오 정보 유사도 행렬은 각각 $I \times I$의 차원을 갖는다. 따라서 아이템 간의 유사도 행렬과 아이템 시나리오 정보 유사도 행렬은 아이템 정보 유사도 행렬이라는 하나의 행렬로 만들 수 있다.

3.3. 아이템 정보 기반 추천 모델

아이템 정보 기반 추천 모델에서는 MF와 아이템 정보 모델을 혼합하여 사용한다. MF의 item latent factor를 추론할 때 아이템 정보 유사도 행렬 정보를 사용한다. 아이템 정보 유사도 행렬 또한 MF와 같이 두 행렬로 분해할 수 있다. 아이템 정보 기반 추천 모델에서는 MF의 item latent factor와 아이템 정보 모델의 item latent factor를 공유하여 아이템 정보 기반 추천 모델을 학습 시킨다. 이 모델은 아래와 같이 표현된다.

$$L = \sum_{u,i} c_{ui}(y_{ui} - \theta_u \beta_i)^2 + \sum_{d_{ij}} (d_{ij} - \beta_i \gamma_j)^2 + \lambda_\theta \sum_u \| \theta_u \|_2^2 + \lambda_\beta \sum_i \| \beta_i \|_2^2 + \lambda_\gamma \sum_j \| \gamma_j \|_2^2 \quad (3)$$

이 수식에서 중요한 점은 item latent factor β_i를 MF와 아이템 정보 모델에서 공유한다는 것이다. 각 파라메터에 대한 최적의 해를 찾을 때, MF와 아이템 정보 모델에 대한 파라메터가 같이 학습된다. 파라메터를 추론하는 dawn liang이 제안한 학습 방법을 사용하여 모델을 학습시킨다[6].

4. 실험

4.1 데이터 셋

본 논문에서 사용한 데이터 셋은 Movielens의 20M 데이터 셋을 사용한다. 이 데이터 셋은 13만 8천명의 사용자와 2만 7천개의 영화에 대한 2천만 개의 평가와 46만 5천개의 태그 정보로 이루어져 있다. 그 중 우리는 사용자, 아이템, 평가 데이터를 사용한다.

아이템에 대한 시나리오 정보는 Movielens의 아이템 아이디를 이용하여 IMDB에서 시나리오 정보를 수집하였다. 전체 아이템 중 시나리오 정보를 수집한 아이템은 11,556개이다.

4.2 실험 및 평가

본 논문에서는 순위 기반의 평가 척도인 mean average precision(MAP) 방식을 사용하였다. MAP은 사용자의 평균 정확도(average precision)을 계산하고, 각 사용자의 평균 정확도의 평균을 구하는 것이다.

$$AP = \sum_{k=1}^{n} P(k) / \min(m, n) \quad (4)$$

m은 실제 아이템에 대한 순위를 나타내고, n은 추천한

아이템 중 실제 소비한 아이템의 수를 나타낸다. P는 정확도를 의미한다.

<표 2> 알고리즘의 평가 결과

시나리오 정보	아이템 정보
0.0469	0.0510

본 논문에서는 추천하는 아이템의 수를 100개로 하여 100개의 순위에 대해서 실험을 진행하였다. 아이템에 대한 시나리오 정보만을 추가하여 실험한 것과 모든 아이템에 대한 정보를 고려하여 실험한 결과, 아이템 정보를 고려하여 실험한 결과가 더 좋게 나왔다.

5. 결론 및 향후 연구 방향

본 논문에서는 아이템 간의 유사도 정보와 아이템의 시나리오 유사도 정보를 사용하여 아이템 정보를 모델링하였으며, 아이템 정보를 기존의 MF에 추가하여 아이템 정보를 기반의 추천 시스템을 제안하고 실험하였다. 아이템의 시나리오 정보를 사용함으로 아이템에 대한 사용자의 선호도 행렬의 비어있는 부분을 보강할 수 있다.

향후 연구 방향으로는 아이템에 대한 정보를 세밀하게 학습시키기 위해 딥러닝을 통해서 아이템을 임베딩하는 것여 추가하는 것과 시계열 정보를 이용하여 아이템에 대한 순서 정보에 대해서 고려하는 것이다.

Acknowledgment

“이 논문은 2016년도 정부 (미래창조과학부) 의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. R1610941).”

참고문헌

[1] Yehuda Koren, Robert Bell, and Chris Volinsky, “Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37, 2009

[2] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. “Recommender systems survey”, Knowledge-Based Systems 46, 109 - 132. 2013

[3] D. Agarwal and B.-C. Chen. “Regression-based latent factor models”, In Proceedings of the 15th ACM SIGKDD, 19 - 28, 2009.

[4] P. K. Gopalan, L. Charlin, and D. Blei. “Content-based recommendations with Poisson factorization”, In Advances in Neural Information Processing Systems, 3176 - 3184, 2014.

[5] H. Shan and A. Banerjee. “Generalized probabilistic matrix factorizations for collaborative filtering” In Data Mining (ICDM), 2010 IEEE 10th International Conference on, 1025 - 1030. IEEE, 2010.

[6] C. Wang and D. Blei. “Collaborative topic modeling for recommending scientific articles”, In Knowledge Discovery and Data Mining, 2011.

[7] Dawen Liang, Jaan Altossar, “Factorization Meets the Item Embedding: Regularizing Matrix Factorization with Item Co-occurrence, RecSys ’16, 15-19, 2016,