추천 시스템에서의 선형 모델과 딥러닝 모델의 데이터 크기에 따른 성능 비교 연구

성다훈[1], 임유진[2]
[1]숙명여자대학교 IT공학과 석사과정
[2]숙명여자대학교 인공지능공학부 교수
ekgns324@sookmyung.ac.kr, yujin91@sookmyung.ac.kr

A Study Comparing the Performance of Linear and Deep Learning Models in Recommender Systems as a Function of Data Size

Da-Hun Seong[1], Yujin Lim[2]
[1]Dept. of Information Technology Engineering, Sookmyung Women's University
[2]Div. of Artificial Intelligence Engineering, Sookmyung Women's University

요 약

추천 시스템을 통해 사용자의 만족도를 높여 매출 증대까지 기대할 수 있기에, 추천 시스템은 과거부터 활발하게 연구되어 왔다. 추천 시스템은 크게 선형 모델과 비선형 모델로 구분할 수 있는데, 각 모델이 주로 독자적으로 연구되어 통합된 성능 결과를 명확히 알 수 없는 경우가 많아, 두 모델 간 특성 차이를 명확히 파악하여 추천 상황에서 적합한 모델을 선택하기 어려운 문제가 있다. 따라서 본 연구에서는 선형 모델과 비선형 모델을 같은 데이터와 같은 환경, 같은 성능평가 지표로 실험하여 결과를 비교 및 분석해보고자 한다.

1. 서론

추천 시스템은 사용자의 평점과 같은 명시적 데이터나 클릭, 구매와 같은 암시적 데이터의 행동 정보를 기반으로 사용자의 선호도를 파악하여 사용자에게 상품이나 콘텐츠 추천을 제공한다. 기업이나 여러 서비스에서는 이러한 추천 시스템으로 개인 맞춤형 경험을 제공하여 사용자의 만족도를 높일 뿐만 아니라 매출 증대까지 기대할 수 있다. 그렇기에 추천 시스템은 과거부터 다양한 접근법을 가진 모델이 제안되어 왔는데, 추천 모델을 크게 구분하자면 선형 모델과 비선형 모델로 나눌 수 있다. 선형 모델은 회귀 계수를 선형 결합으로 표현할 수 있는 모델이며, 비선형 모델은 회귀 계수를 선형 결합으로 표현할 수 없는 것으로 대표적으로 딥러닝 모델이 해당된다. 딥러닝 모델은 일반적으로 선형 모델보다 성능이 좋지만, 낮은 차원의 임베딩 공간으로 인해 표현력이 제한될 수 있다[1]. 그렇기에 딥러닝 모델과 비교하여 실험적으로 성능 차이가 있음에도, 실제 환경에서 딥러닝 모델의 대안이 될 수 있는 선형 모델 추천 시스템이 여전히 활발히 연구되고 있다. 그러나 각 모델이 주로 독자적으로 연구되고 있어, 통합된 성능 결과를 명확히 알 수 없는 경우가 많아, 두 모델 간 특성 차이를 명확히 파악하여 추천 상황에서 적합한 모델을 선택하기 어려운 문제가 있다. 따라서 본 연구에서는 선형 모델과 비선형 모델을 같은 데이터와 같은 환경, 같은 성능평가 지표로 실험하여 결과를 비교 및 분석해보고자 한다.

2. 활용 모델

본 연구에서는 선형 모델로 EASER[2]를, 비선형 모델로 BERT4Rec(Bidirectional Encoder Represent-ations from Transformers for sequential Recomm-endation)[3]을 사용하였다. EASER는 최신 선형 추천 모델인 EASE(Embarrassingly Shallow Autoe-ncoders for Sparse Data)[4]를 확장한 모델로, 명시적 입력 특성을 추가하여 다른 선형 모델과 달리 고차(Higher-order) 상호작용과 네거티브(Negative) 상호작용을 고려함으로써 추천 성능을 높인 선형 모델이다. 이 모델은 사용자가 상호작용한 아이템 쌍 사이의 관계를 고려하여 사용자의 이전 상호작용 기록에서 두 아이템 이상의 조합이 더 높은 관련성을

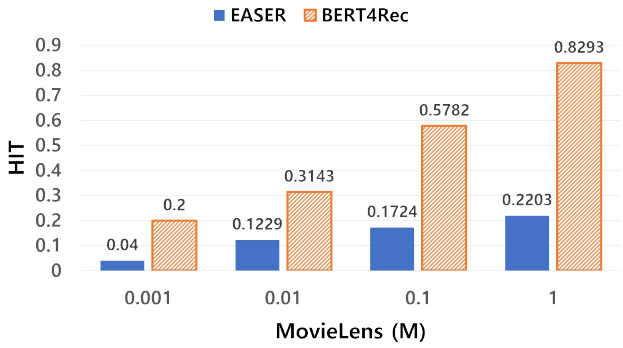
갖는 경우를 식별하고, 이를 통해 보다 복잡한 사용자 행동 양식과 선호도를 포착할 수 있다. 또한 이 모델은 네거티브 상호작용을 허용하여 사용자의 부정적인 선호도나 관심 부족을 표현하는 데 유용하다. 사용한 MovieLens[5] 데이터에는 평점 점수를 통해 영화에 대한 사용자의 부정적 선호도를 파악할 수 있어 이러한 모델이 적합하다. BERT4Rec 또한 비선형 모델 중, 최신 딥러닝 추천 모델로, NLP(Nat-ural Language Processing) 분야에서 사용되는 양방향 자기 주의(Bidirectional Self-attention) 메커니즘인 BERT(Bidirectional Encoder Representations from Transformers)[6]를 사용하여 사용자의 행동 시퀀스를 모델링한다. 이때 언어 모델에서 일반적으로 사용되는 기술인 클로즈 작업(Cloze Task)을 사용하여, 시퀀스의 임의 항목을 마스킹하고 모델이 문맥을 기반으로 이러한 마스킹된 항목을 예측한다. 이는 다른 딥러닝 모델이 갖는 단방향 문제인, 각 항목이 과거의 항목으로부터만 제한된 정보를 인코딩한다는 문제와, 사용자의 행동이 일률적인 순서만을 따르는 것은 아니라는 문제를 해결한다. BERT4Rec와 같은 양방향 순차 추천 모델은 실험에 사용한 데이터셋처럼 사용자의 영화 평점 정보가 시간순으로 기록되어 있는 경우에 유용하다.

3. 실험 설계

본 연구는 추천 시스템에서 데이터의 크기에 따른 선형 모델인 EASER와 비선형 모델인 BERT4Rec의 성능 차이를 비교하기 위해서 MovieLens 데이터셋을 이용하였다. MovieLens 데이터셋은 추천 시스템에서 많이 활용되는 데이터셋으로, 사용자의 영화 평점 데이터가 포함되어있다. 데이터셋의 크기는 0.001, 0.01, 0.1, 1(백만)개로 4가지의 경우로 나누었으며, 성능평가 지표로는 HIT(Hit Rate)와 NDCG(Normalized Discounted Cumulative Gain)를 사용하였다. HIT@K와 NDCG@K는 랭킹 기반 추천 시스템에서 많이 쓰이는 평가지표로, HIT는 전체 사용자 수 대비 적중한 사용자 수를 나타내며, NDCG는 추천 순서를 포함하여 가장 이상적인 추천 조합 대비 현재 모델의 추천 리스트가 얼마나 좋은지를 나타내는 지표이다. HIT는 NDCG와 달리, 추천 랭킹을 고려하지 않는다는 단점이 있으나, 사용자가 실제로 선택한 항목이 추천 목록에 포함되어 있는 비율을 알려주기 때문에 NDCG와 함께 사용되고 있다. 두 성능평가 지표는 모두 0에서 1사이의 값을 가진다. 추천 아이템의 수를 나타내는 K는 10으로 설정하여 NDCG@10과 HIT@10일 때의 성능 결과를 확인하였다.

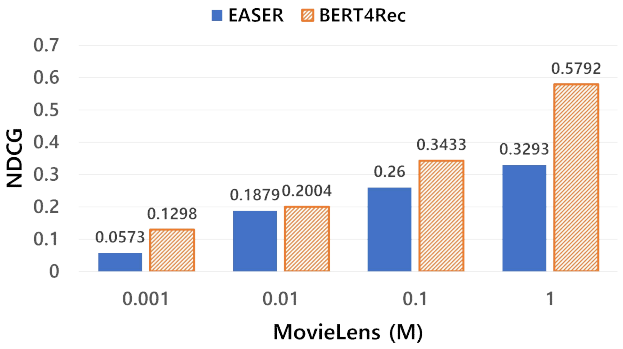
4. 실험 결과

실험 결과로는 데이터의 크기에 따른 선형 모델인 EASER와 비선형 모델인 BERT4Rec의 HIT@10와 NDCG@10의 성능 결과를 살펴볼 것이다. 먼저 HIT@10의 결과는 (그림 1)과 같다.



(그림 1) HIT@10 성능 평가 결과

두 모델 모두 데이터가 증가할수록 성능 점수도 향상되는데, EASER는 모든 데이터 크기에서 전반적으로 성능 점수가 낮게 나왔으나, BERT4Rec는 데이터 증강에 따라 성능 점수가 큰 폭으로 향상되며, 특히 데이터가 1M일 때 성능 점수가 0.8293으로 우수한 성능 결과를 보인다. 이는 이 모델이 약 80%의 사용자에게 사용자가 실제로 선택한 아이템을 추천했다는 것을 의미한다. 또한 BERT4Rec는 EASER에 비하여 각 데이터 크기별로 약 4배 정도의 큰 차이로 높은 성능을 보여주는데, HIT는 전체 사용자에 대한 추천 성능 결과를 보여주므로, 이러한 성능 차이는 딥러닝 모델이 선형 모델에 비해 사용자의 특성에 대한 일반화 능력이 큰 차이로 높다는 것을 알 수 있다.



(그림 2) NDCG@10 성능 평가 결과

(그림2)를 보면, NDCG 결과는 HIT의 결과와 마찬가지로, 두 모델 모두 데이터가 증가할수록 성능도 향상되는데, BERT4Rec가 EASER보다 성능 점수가 높기는 하나 HIT 결과와 달리, 데이터가 0.1M까지는 순위 추천 결과에 있어서, 두 모델의 성능 차이가 크지 않다는 것을 알 수 있다. 이는 초기 데이터 세트 크기에서는 BERT4Rec의 더 복잡한 딥러닝 기반 구조가 EASER보다 상당한 이점을 제공하지 않을 수 있다는 것을 의미한다.

EASER와 BERT4Rec의 HIT와 NDCG의 결과를 종합적으로 살펴 보았을 때, 두 모델 모두 데이터의 크기가 커질수록 각 성능 점수가 커지는 것을 확인할 수 있다. 그러나 BERT4Rec은 EASER에 비하여, 데이터가 늘어날수록 큰 폭으로 성능 향상을 보이는데, 이는 BERT4Rec와 같은 딥러닝 모델이 선형 모델에 비하여 데이터의 크기에 의존성이 크다는 것을 알려준다. 특히 데이터가 1M으로 방대할 때 BERT4Rec의 각 지표에서 성능 점수가 두드러지는데, 이는 데이터가 많은 상황에서 비선형 모델이 선형 모델에 비해 높은 성능을 보인다는 것을 의미한다. 또한 전체 실험에서 BERT4Rec가 EASER보다 높은 성능 점수를 보였다. 이는 딥러닝 모델이 비선형 함수를 통해 사용자와 아이템의 고차 상호작용을 학습함으로써, 사용자의 숨겨진 선호도나 복잡한 관계를 학습했기 때문으로 보인다. 일반적으로 NDCG가 HIT에 비하여 더 정교한 성능 지표이므로 점수가 더 낮게 나오는 경향이 있는데, EASER의 경우 HIT보다 NDCG일 때 점수가 더 높게 나오는 것을 관찰할 수 있었다. 이는 EASER 모델이 일부 사용자의 매우 높은 만족도를 달성할 수 있지만, 다수 사용자에게 만족도를 제공하지 못한다는 것을 의미한다. 이는 선형 모델이 비선형 모델과 달리 다양한 패턴을 파악할 수 없다는 한계를 보여주는 것으로 보인다.

5. 결론

실험 결과를 통해 비선형 모델인 BERT4Rec가 선형 모델인 EASER보다 전반적인 성능이 높게 나왔으며, 특히 일반화 능력이 중요한 경우, 데이터셋의 크기에 관계없이 딥러닝 모델이 선형 모델에 비하여 큰 차이로 높은 성능을 보이므로 딥러닝 모델을 사용하는 것이 유리함을 알 수 있다. 그러나 순위 추천에 있어서는 특정 데이터 크기 이하에서는 EASER와 BERT4Rec 사이의 성능 차이가 크지 않다는 것을 확인했다. 이는 추천 시스템의 선택에 있어 데이터 세트의 크기가 큰 영향을 미칠 수 있음을 시사한다. 또한 딥러닝 모델은 데이터의 크기에 의존성이 있고 모델 복잡성이 크지만, 선형 모델은 비교적 견고하고 계산적으로 효율적이라는 장점을 갖기 때문에, 추천 모델을 선택할 때 데이터셋의 크기와 문제 상황에 따라 선형 모델과 비선형 모델을 비교한 성능 검사가 우선되어야 할 것이다. 본 연구는 추천 상황에서 선형 모델과 비선형 모델을 선택할 때 유용한 통찰을 제공할 것으로 기대한다.

본 연구에서는 선형 모델과 비선형 모델로 각각 하나의 모델을 대표하여 비교하였으므로, 향후 연구에서는 다양한 선형 모델과 비선형 모델을 추가하여 성능 비교 연구를 진행할 계획이다. 또한 성능 비교 실험에서 데이터의 크기 외에도 풀고자 하는 문제의 성격, 데이터의 특성, 해석 가능성의 중요성 등 여러 요인이 있기 때문에 실험에 사용한 영화 데이터와 다른 특성을 갖는 다양한 데이터셋에서의 성능 비교 연구를 진행하고자 한다.

사사문구

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0 사업의 연구결과로 수행되었음” (IITP-2024-RS-2022-00156299)

참고문헌

[1] A. Beutel, P. Covington, S. Jain, C. Xu, J. Li, V. Gatto and E.H. Chi, “Latent Cross: Making Use of Context in Recurrent Recommender Systems,” Proceedings of in ACM Conference on Web Search and Data Mining (WSDM), pp. 46-54, Feb. 5-9, Marina Del Rey, CA, USA, 2018.

[2] H. Steck and D. Liang, “Negative Interactions

for Improved Collaborative Filtering: Don't Go Deeper, Go Higher," Proceedings of the 15th ACM Conference on Recommender Systems (RecSys), pp. 34 - 43, Sep. 27-Oct. 1, Amsterdam, Netherlands, 2021.

[3] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou and P. Jiang. "BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer," Proceedings of in the 28th ACM International Conference on Information and Knowledge Management (CIKM), pp. 1441-1450, Nov. 3 - 7, Beijing, China, 2019.

[4] H. Steck, "Embarrassingly Shallow Autoencoders for Sparse Data," Proceedings of the World Wide Web Conference (WWW), pp. 3251 - 3257, May. 13 - 17, CA, USA, 2019.

[5] Kaggle의 MovieLens 데이터셋, https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset?select=genome_tags.csv

[6] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of ArXiv Preprint, 2018, arXiv:1810.04805.