뉴스 추천 시스템에서의 제목 인덱싱의 활용 가능성 분석

김준표[1], 김태호[2], 김상욱[3]
[1]한양대학교 컴퓨터소프트웨어학과 석사과정
[2]한양대학교 컴퓨터소프트웨어학과 박사과정
[3]한양대학교 컴퓨터소프트웨어학과 교수

pyo9912@agape.hanyang.ac.kr, hirooms2@agape.hanyang.ac.kr, wook@agape.hanyang.ac.kr

Analysis of the feasibility of using title-id indexing in a news recommendation system

Jun-Pyo Kim[1], Tae-Ho Kim[2], Sang-Wook Kim[3]
[1]Dept. of Computer Software Engineering, Hanyang University
[2]Dept. of Computer Software Engineering, Hanyang University
[3]Dept. of Computer Software Engineering, Hanyang University

요 약

현재까지 연구되었던 뉴스 추천 시스템은 일반적으로 뉴스 제목, 뉴스 본문, 카테고리 정보 등의 텍스트 정보를 기반으로 사용자에게 맞춤 뉴스를 추천해주는 방식으로 동작한다. 구체적으로는 뉴스의 텍스트 정보를 통해 뉴스를 표현하는 임베딩 벡터를 생성하여 사용자 맞춤 뉴스를 추천하는 task-specific 한 아키텍처를 기반으로 동작한다. 기존 연구에서는 task-specific 아키텍처 내의 뉴스의 임베딩 벡터를 생성하는 과정에서 BERT 와 같은 언어모델을 이용하여 텍스트 정보를 더 잘 반영하고자 했다. 본 연구에서는 기존의 구조와 다르게, 뉴스 제목 인덱싱을 통해 전체 뉴스 추천 시스템에서의 언어모델을 충분히 활용할 수 있는 방식을 제안하고자 한다.

1. 서론

추천 시스템은 지난 수십년간 사람들의 일상에서 큰 역할을 수행하고 있었으며, 최근에는 여러 연구 및 실생활 서비스에서도 더 종합적이고 넓은 스펙트럼을 가지는 시스템을 개발하고자 한다.

초기의 추천 시스템은 로지스틱 회귀 또는 협업 필터링[1-3]과 같은 방식을 이용하여, 모델이 사용자-아이템 상호작용 정보를 학습하여 사용자의 클릭 패턴을 예측할 수 있도록 하였다. 뿐만 아니라 사용자 프로필이나 아이템의 메타 데이터와 같은 contextual features 를 이용하는 방법들도 연구되었다.

이후 딥러닝 방식이 적용되면서 아이템의 features 를 더 잘 다루도록 하는 모델들이 연구되었다[4-6]. 뉴스 추천 시스템에서는 사용자가 클릭한 history news 들을 기반으로 candidate news 의 클릭 여부를 판단하는데, 이때 뉴스의 제목, 본문, 카테고리 등 텍스트에 내재된 features 를 추출하기 위해 딥러닝을 이용하는 방식이 제안되었다. 최근 제안된 방식들은 언어모델을 이용하여 뉴스를 더 효과적으로 표현하였고, 언어모델의 힘을 통해 추천 성능을 높인 것을 확인할 수 있었다.

이처럼 언어모델은 뉴스 추천 시스템의 성능을 높이는데 기여한다. 그러나 현재까지 제안된 대부분의 뉴스 추천 방식들은 언어모델을 뉴스의 features 를 추출하는 단계에서만 사용하고 있다. 일반적인 뉴스 추천 시스템은 뉴스를 표현하는 뉴스 인코더와 사용자 정보를 표현하는 사용자 인코더 두 단계로 구성되는데, 뉴스 인코더에서는 언어모델을 사용하는 반면 사용자 인코더는 단순한 딥러닝 연산을 사용하고 있다.

이렇게 두 인코더가 구분되는 구조는 언어모델의 힘을 충분히 활용하지 못하게 한다는 한계점이 존재한다. 왜냐하면 언어모델이 pre-train 된 방식과 사용자 맞춤 뉴스를 추천하는 downstream task 사이의 차이가 존재하기 때문이다.

또한 현재 언어모델이 점차 커지면서 언어모델을 이용하는 패러다임이 바뀌고 있는데, 하나의 모델을 이용하여 통합된 task 를 수행하는 패러다임이 새로 생기면서 transfer learning 에 대한 연구도 활발히 진행되고 있다[7-9].

따라서 전체 시스템에서 언어모델의 힘을 충분히 이용하기 위해서는 사용자의 history news 를 입력 받는 과정부터 candidate news 의 클릭 여부를 출력하는 과정까지 언어모델을 이용하여 언어모델이 pre-train 된 방식과 동일한 방식으로 수행해야 한다.

이를 효과적으로 하기 위한 수단으로 본 논문에서는 뉴스 제목을 아이디로 인덱싱하는 방법을 제안하며, 뉴스 제목 인덱싱의 효과를 알아보고자 한다.

2. 모델 아키텍처

전체 모델 아키텍처 관련하여 본 연구에서는 Transformers 기반의 인코더-디코더 구조의 언어모델을 이용하고자 한다. 기존의 연구에서 뉴스의 임베딩을 구하기 위해 인코더 구조의 언어모델을 이용한 것과 달리 인코더-디코더 구조의 언어모델을 이용하게 되면 입력 시퀀스를 처리한 결과를 임베딩 값이 아닌 자연어 형태로 출력할 수 있게 된다. 그리고 이러한 형태의 출력은 언어모델이 pre-train 된 방식인 Masked Language Modeling 과 동일한 형태의 task 를 수행할 수 있도록 돕는다.

구체적으로 인코더-디코더 구조를 통해 본 연구에서 목표로 하는 task 는 뉴스 제목을 아이디로 기억하는 것이다. 그래서 입력 토큰 시퀀스 $x = [x_1, \cdots, x_n]$ 을 입력 받아 인코더 $\mathcal{E}(\cdot)$ 을 거쳐 입력 토큰 시퀀스의 맥락 정보를 반영한 contextualized textual representations $t = [t_1, \cdots, t_n]$ 을 생성하도록 한다. 이후 디코더 $\mathcal{D}(\cdot)$에서는 이전까지 출력인 $y_{<i}$ 및 인코더의 출력인 t를 참고하여 출력 토큰 y를 생성하도록 한다. 즉, $P_\theta(y_j \mid y_{<j}, x) = \mathcal{D}(y_{<j}, t)$ 을 뜻하게 되며, 본 아키텍처에서는 label 토큰 y 에 대한 log-likelihood 를 줄이도록 모델 파라미터 θ 을 학습한다.

$$L_\theta = -\sum_{j=1}^{|y|} \log P_\theta\,(y_j | y_{<j}, x)$$

위와 같은 방식으로 모델이 입력 시퀀스로 주어진 뉴스의 텍스트 정보를 보고 뉴스 아이디를 맞게 출력하도록 학습시킨다.

3. 실험

우리의 방식을 학습하고 평가하기 위해, 실생활 뉴스 데이터 셋인 MIND 데이터셋을 이용한다. 이는 Microsoft News 에 대한 사용자 행동패턴 로그가 담긴 데이터셋으로 MIND-large 와 MIND-small 두 버전이 존재한다. 그 중 우리는 MIND-large 버전을 이용하고 있으며, <표 1>에 요약된 실세계 데이터셋을 사용한다.

<표 1> 데이터셋 통계

	MIND-large
# News	161,013
# Categories	20
# Impressions	15,777,377
# Clicks	24,155,470

본 실험은 MIND-large 데이터셋에 존재하는 뉴스들에 대하여 랜덤한 수의 샘플을 뽑아서 뉴스의 제목, 본문의 내용을 보고 뉴스의 아이디를 맞히는 정도를 확인하고 있다. 우리는 인코더-디코더 구조의 언어모델 중 fine-tuning 이 가능한 T5 모델[10]을 이용하여 실험을 진행하였으며, <표 2>와 <표 3>에는 각각 뉴스의 제목을 보고 아이디를 맞힐 때의 정확도, 뉴스의 제목과 본문을 보고 아이디를 맞힐 때의 정확도를 측정한 결과이다.

<표 2> 뉴스 제목 정보를 기반으로 아이디 추론 시 정확도

Model	# Sample	Accuracy
T5	1,000	99.6%
T5	2,000	99.5%
T5	5,000	99.7%

<표 3> 뉴스 제목 및 본문 정보를 기반으로 아이디 추론 시 정확도

Model	# Sample	Accuracy
T5	1,000	91.1%
T5	2,000	92.7%
T5	5,000	80.7%

실험 결과 뉴스의 제목 정보를 기반으로 아이디를 추론하는 것이 뉴스의 제목 및 본문 정보를 기반으로 아이디를 추론하는 것보다 효과적임은 알 수 있었다. 특히, 샘플 수가 5,000 개인 상황에서 정확도 차이가 두드러지게 나타난 것을 확인할 수 있었다.

4. 결론 및 향후 연구

본 논문은 뉴스 추천 시스템이 언어모델의 힘을 충분히 이용하도록 하는 패러다임을 제시하면서, 그 과정 속에서 시스템의 입력으로 뉴스를 아이디로 제공할 수 있는지 여부에 대해 알아보았다.

실험 결과에서도 보이듯 일정 수의 샘플에서 뉴스 제목을 보고 아이디를 맞히는 작업에서 어느정도 높은 정확도를 보이는 것을 확인했으며, 뉴스 아이디를 인덱싱하는 것이 효과적이다는 결론을 도출하였다.

향후 연구에서는 이 실험 결과를 바탕으로 실제 사용자 맞춤 뉴스 추천 작업까지 수행하는 것을 계획하고 있으며, 언어모델의 힘을 온전히 가져가는 패러다임의 효과성을 입증하고자 한다.

5. 사사

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이고 (No.2022-0-00352), 2018 년도 정부(과학기

술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이고(No.2018R1A5A7059549), 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155586, 실세계의 다양한 다운스트림 태스크를 위한 고성능 빅 하이퍼그래프 마이닝 플랫폼 개발(SW 스타랩))

참고문헌

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

[2] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet computing 7, 1 (2003).

[3] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work. 175–186

[4] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural collaborative reasoning. In Proceedings of the Web Conference 2021. 1516–1527.

[5] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In Proceedings of the 1st workshop on deep learning for recommender systems. 7–10.

[6] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence. 1725–1731.

[7] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 4947–4957

[8] Kyuyong Shin, Hanock Kwak, Kyung-Min Kim, Minkyu Kim, Young-Jin Park, Jisu Jeong, and Seungjae Jung. 2021. One4all User Representation for Recommender Systems in E-commerce. arXiv preprint arXiv:2106.00573 (2021).

[9] Kyuyong Shin, Hanock Kwak, Kyung-Min Kim, Su Young Kim, and Max Nihlen Ramstrom. 2021. Scaling Law for Recommendation Models: Towards Generalpurpose User Representations. arXiv preprint arXiv:2111.11294 (2021).

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html