Towards Efficiency and Sparsity: Exploiting Personalized Terms for News Recommendation

Anonymous Author(s)

ABSTRACT

Some terms in a user's browsed news reveal her interests from the most fine-grained level. Capturing these **personalized terms** yields a highly compact and accurate user profile, significantly reducing the cost of using effective yet expensive matching models (e.g.BERT). The sparse terms bring interpretability to recommenders: the candidate generation process can be aligned to adhoc retrieval if we consider the personalized terms as query and the whole news set as document collection, where fast retrieval techniques can be employed. Due to lack of ground truth labels, we empower the model to learn to select representative terms, gaining a consistent improvement over heuristic selection baselines.

1 INTRODUCTION

2 RELATED WORK

In this section, we review the related work of news recommendation, feature selection and candidate recall.

2.1 News Recommendation

News recommendation has been widely explored for decades. Traditional collaborative filtering methods [5, 12] hash similar users into the same group by LSH before recommending. They either employ Matrix Factorization [10] to gather users. MF decomposes the user-item matrix to map users and items into the same latent space, where the inner product between their vectors captures the interaction score. Unlike movies or products, enoumous news is spawn every second, outstriping the increase of users, which makes the user-item matrix especially sparse. Factorization Machine [15] then introduces real-valued features to MF to alleviate the sparsity, but it requires manual features which is time and labor demanding.

Content-based news recommendation then emerges to address the above issues. Early works of content-based methods still rely on some manual features such as trend [13], geographics [11], and demographics [4] to model news and users. In the recent ten years, deep learning shows an unlimited potential and prevails in the task of representation learning [3]. So more and more works are underway to design equisite structures to learn representation of news and users directly from raw texts and browsing history respectively, taking the dot product between them as the click probability. Wu et al. [1, 17–22] proposes effective models that employs CNN, RNN, multi-head attention, and personalized attention to represent news and users. External information such as knowledge [16] and user-item biparititie graph [9, 22] are also incorporated to enhance representations.

2.2 Efficient Transformers

More recently, large-scale pretrained language models (PLM in short) e.g. BERT [6] demonstrates impressive improvements over shallow and light-weighted neural models in NLP field. However, it is non-trivial to implement PLMs in the news recommendation scenario: the quadratic complexity of self-attention w.r.t. the input sequence length poses an intense challenge to speed up encoding users with dozens of historical articles that may contain thousands of words in total. Researchers make a lot of efforts to lessen this problem. SpeedyFeed [24] deduplicates user's historical news and candidate news in a batch, significantly speeding up the training process. Apart from optimizing training scheme in industry, more research modifies the self-attention layer to reduce complexity. For example, Longfomer [2] sparsifies the full self-attention to a sliding and dialated pattern, which reduces the complexity to linear w.r.t. the input length; Fastformer achieves the same result by additive attention and element-wise production. It also reaches a new state-of-the-art performance on MIND [23], a large-scale dataset in news recommendation.

Another line of research followes a feature selection setting that prunes the input for expensive interaction. Hofstätter et al. [8] restricts BERT to only inspect top K important passages per document. It splits the selection and ranking stage, where the former is trained in teacher-student paradiagm by the pseudo labels produced by a BERT, and the latter only scores the selected passages. Using the similar cascading setting, Pi et al. [14] extracts fewer valueble items from user history to feed into final ranking. However, cascading achitechture requires labels for each stage, which is prohibitive in our scenario because there is no ground-truth indicating terms that the user really favors. Gallagher et al. [7] explores framework for jointly optimizing cascade search, but it is not practicable in a BERT-style model since it relies on specified empirical risk rather than ranking loss. The first application of sparse selection in news recommendation is SFI [**?**]. It sparsely and automatically selects important historical news for effective candidate-aware interaction, guaranteeing the efficiency and effectiveness of the model. Despite its improvements, SFI executes selection at new level, possibly bringing high bias to the final ranking. It also neglects PLMs to promote performance.

In our work, we select the browsing history at word-level to keep more fine-grained and comprehensive information instead of pruning the historical news set. With only a handful of personlaized terms, we apply PLMs to promote the effectiveness of the recommender to a new level at competitive speed.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

2.3 Sparse Recall

As a bonus, personalized terms generated by our model can transform the entire user history to a rather short query, which can greatly facilitate candidate recall. Here we briefly introduce some retrieval techniques in IR.

REFERENCES

[1] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *ACL,2019*. Association for Computational Linguistics, 336–345.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR* abs/2004.05150 (2020). arXiv:2004.05150 https://arxiv.org/abs/2004.05150

[3] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

[4] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *DLRS@RecSys, 2016*. ACM, 7–10.

[5] Abhinandan Das, Mayur Datar, Ashutosh Garg, and Shyamsundar Rajaram. 2007. Google news personalization: scalable online collaborative filtering. In *WWW,2010*. ACM, 271–280.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT,2019*. Association for Computational Linguistics, 4171–4186.

[7] Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J Shane Culpepper. 2019. Joint optimization of cascade ranking models. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 15–23.

[8] Sebastian Hofstätter, Bhaskar Mitra, Hamed Zamani, Nick Craswell, and Allan Hanbury. 2021. Intra-Document Cascading: Learning to Select Passages for Neural Document Ranking. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1349–1358. https://doi.org/10.1145/3404835.3462889

[9] Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020. Graph Neural News Recommendation with Unsupervised Preference Disentanglement. In *ACL,2020*. Association for Computational Linguistics, 4255–4264.

[10] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.

[11] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *WWW, 2010*. ACM, 661–670.

[12] Lei Li, Dingding Wang, Tao Li, Daniel Knox, and Balaji Padmanabhan. 2011. SCENE: a scalable two-stage personalized news recommendation system. In *SIGIR,2011*. ACM, 125–134.

[13] Jiahui Liu, Peter Dolan, and Elin Rønby Pedersen. 2010. Personalized news recommendation based on click behavior. In *IUI,2010*. ACM, 31–40.

[14] Qi Pi, Guorui Zhou, Yujing Zhang, Zhe Wang, Lejian Ren, Ying Fan, Xiaoqiang Zhu, and Kun Gai. 2020. Search-based User Interest Modeling with Lifelong Sequential Behavior Data for Click-Through Rate Prediction. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 2685–2692. https://doi.org/10.1145/3340531.3412744

[15] Steffen Rendle. 2010. Factorization Machines. In *ICDM,2010*. IEEE Computer Society, 995–1000.

[16] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-Aware Network for News Recommendation. In *WWW,2018*. ACM, 1835–1844.

[17] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-View Learning. In *IJCAI,2019*. ijcai.org, 3863–3869.

[18] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *SIGKDD,2019*. ACM, 2576–2584.

[19] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Topic-Aware News Representation. In *ACL,2019*. Association for Computational Linguistics, 1154–1159.

[20] Chuhan Wu, Fangzhao Wu, Mingxiao An, Tao Qi, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Heterogeneous User Behavior. In *EMNLP-IJCNLP,2019*. Association for Computational Linguistics, 4873–4882.

[21] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *EMNLP-IJCNLP,2019*. Association for Computational Linguistics, 6388–6393.

[22] Chuhan Wu, Fangzhao Wu, Tao Qi, Suyu Ge, Yongfeng Huang, and Xing Xie. 2019. Reviews Meet Graphs: Enhancing User and Item Representations for Recommendation with Hierarchical Attentive Graph Neural Network. In *EMNLP-IJCNLP,2019*. Association for Computational Linguistics, 4883–4892.

[23] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *ACL,2020*. Association for Computational Linguistics, 3597–3606.

[24] Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, and Xing Xie. 2021. Training Large-Scale News Recommenders with Pretrained Language Models in the Loop. *arXiv preprint arXiv:2102.09268* (2021).