Towards Efficiency and Sparsity: Exploiting Personalized Terms for News Recommendation

Anonymous Author(s)

1 INTRODUCTION

With the ever growing data and improving technologies, more and more news overwhelms people with its countless topics and events. In this information torrent, an individual is nearly impossible to pick out small drops that she's interested in. Luckily, online news platforms e.g. MSN[1] serve that function for each user, significantly alleviating the information overload problem [15]. The core technology of such platforms is personlaized news recommendation.

News recommendation comprises of two major steps: recall and rerank: the former recalls a small subset of news that is relevant to the user's preference from the entire news collection; the latter reranks this subset in the descending order of click probability. In both stages, efficiency is the top concern [18, 20**?**].

The mainstream candidate recall methods rely on similarity search: they represent users and news articles as vectors in the same semantic space, where the top K nearest neighbors of a user representation are retrieved as relevant candidates. In IR, the sparse counterpart against this dense method, which retrieves documents with exact match of the query by inverted index, is even more efficient [**?**]. It also provides more interpretable and controlled patterns that appeals to the industry [11]. However, lacking of an explicit query hinders the implementation of sparse recall. Early works labels each news with some pre-defined tags or keywords, then generate a user profile by collecting tags of her browsed news. However, it's always difficult to design tags at proper granularity that is precise enough to match relevant candidates and low-cost enough to conduct and maintain. Another concern of these methods is personalization. Different users may attend to different aspects of news [**?**], however, the keywords of one piece of news is the same regardless of users. Recent methods exploit heuristics [26] or neural networks [] to extract keywords from historical news text, which improves precision and avoids laborous engineering. Yet the gap to personalization remains.

When it comes to rerank stage, the two-tower structure is the most popular. It encodes each news article and each user into a compact vector respectively, then computes the similarity between them as the click probability. Under this setting, various news/user encoder has been explored, such as RNN [**?**], CNN [28], and a number of attentive methods [32**? ?**]. Recently, Wu et al. [34] integrates PLMs to be the news encoder and validates their effectiveness. Despite improving accuracy, PLMs consume much more time and space when encoding due to their deep and large-scale structure. This problem is even worse in news recommendation: the model must encode all of the user's historical news and every candidate news for a single impression, which may contain thousands of words in total. In consequence, existing methods are forced to encode each historical news then feed a pooled representation to the user encoder. In this workflow, fine-grained interactions across different news are lost, which is proved to be useful in IR. [**?**]

[1] https://www.msn.com/en-us/news

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 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 [8, 17] hash similar users into the same group by LSH before recommending. They either employ Matrix Factorization [14] to gather similar 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, and makes the user-item matrix especially sparse. Factorization Machine [25] 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 [19], geographics [16], and demographics [5] to model news and users. In the recent ten years, deep learning techniques prosper and greatly promote the progress of news recommendation systems. 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. This *two tower* workflow appeals to industry since encoding of news and user can be done offline, greatly improving model efficiency. Under the two tower setting, Wu et al. [1, 28–33] proposes effective models that employs CNN, RNN, multi-head attention, and personalized attention to represent news and users. External information such as knowledge [27] and user-item biparititie graph [13, 33] are also incorporated to enhance representations.

2.2 Efficient Transformers

More recently, large-scale pretrained language models (PLM in short) e.g. BERT [9] demonstrates impressive improvements over shallow and light-weighted neural models in NLP field. Though a previous work [34] integrates PLMs to news recommendation and validates their improvements, it is yet non-trivial to efficiently implement them: 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 [36] 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 [3] 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 achieves a new state-of-the-art performance on MIND [35], a large-scale dataset in news recommendation.

Another line of research followes a feature selection intuition that prunes the input before expensive interaction. Hofstätter et al. [12] 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. [24] 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. [10] explores framework for jointly optimizing cascade search, but it is not practical 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 before effective candidate-aware interaction, guaranteeing the efficiency and effectiveness of the model. Despite its improvements, SFI executes selection at new level, possibly losing information hidden in the unselected news for the final ranking. It also neglects PLMs to promote performance due to its one tower limitation.

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 fully capture the interaction within and among historical articles, promoting the effectiveness of the news recommender to a new level at a competitive speed.

2.3 Candidate Recall

The mainstream candidate recall methods rely on similarity search: they project users and items into the same latent space, where the top K nearest neighbors of a user are retrieved as relevant items. In IR, the sparse counterpart of such dense recall, which retrieves documents with exact match of the query, is even more efficient [**?**]. However, there is no explicit query in recommendation, prohibiting the use of sparse recall techniques.

In our work, the generated personalized terms natually compresses the entire user history into a rather short query, with which we can align the candidate recall to ad-hoc retrieval and take advantage of sparse recall techniques. Due to space limitations, we only explore the naive sparse recall with personalized terms in this paper, and leave other advanced methods such as COIL [11] to future works.

Type	**Methods**	**AUC**	**MRR**	**NDCG@5**	**NDCG@10**
Sota	FastFormer	**72.68**	**37.45**	**41.51**	**46.84**
Baselines	Two Tower	71.43	36.16	39.67	45.29
	TES-First	68.00	32.91	36.44	42.86
	TES-BM25	68.00	32.91	36.44	42.86
	TES-Entity	68.00	32.91	36.44	42.86
Ours	TES	69.62	34.30	37.47	43.21

3 EXPERIMENT

3.1 Results

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] Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. SparTerm: Learning Term-based Sparse Representation for Fast Text Retrieval. *CoRR* abs/2010.00768 (2020). arXiv:2010.00768 https://arxiv.org/abs/2010.00768

[3] 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

[4] 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

[5] 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.

[6] Zhuyun Dai and Jamie Callan. 2019. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval. *CoRR* abs/1910.10687 (2019). arXiv:1910.10687 http://arxiv.org/abs/1910.10687

[7] Zhuyun Dai and Jamie Callan. 2020. Context-Aware Document Term Weighting for Ad-Hoc Search. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 1897–1907. https://doi.org/10.1145/3366423.3380258

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

[9] 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.

[10] 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.

[11] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 3030–3042. https://doi.org/10.18653/v1/2021.naacl-main.241

[12] 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

[13] 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.

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

[15] Talia Lavie, Michal Sela, Ilit Oppenheim, Ohad Inbar, and Joachim Meyer. 2010. User attitudes towards news content personalization. *Int. J. Hum. Comput. Stud.* 68, 8 (2010), 483–495.

[16] 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.

[17] 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.

[18] Defu Lian, Haoyu Wang, Zheng Liu, Jianxun Lian, Enhong Chen, and Xing Xie. 2020. LightRec: A Memory and Search-Efficient Recommender System. In *WWW,2020*. ACM / IW3C2, 695–705.

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

[20] Zheng Liu, Yu Xing, Fangzhao Wu, Mingxiao An, and Xing Xie. 2019. Hi-Fi Ark: Deep User Representation via High-Fidelity Archive Network. In *IJCAI,2019*. ijcai.org, 3059–3065.

[21] Antonio Mallia, Omar Khattab, Torsten Suel, and Nicola Tonellotto. 2021. Learning Passage Impacts for Inverted Indexes. 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, 1723–1727. https://doi.org/10.1145/3404835.3463030

[22] Donald Metzler and W. Bruce Croft. 2005. A Markov random field model for term dependencies. In *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005*, Ricardo A. Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait (Eds.). ACM, 472–479. https://doi.org/10.1145/1076034.1076115

[23] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019). arXiv:1904.08375 http://arxiv.org/abs/1904.08375

[24] 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

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

[26] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

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

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering News Recommendation with Pre-trained Language Models. 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, 1652–1656. https://doi.org/10.1145/3404835.3463069

[35] 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.

[36] 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).

[37] Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik G. Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 497–506. https://doi.org/10.1145/3269206.3271800