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 [? ?] hash similar users into the same group by LSH before recommending. Matrix Factorization [?] is also a popular technique to gather users. It decomposes the user-item matrix to map users and items into the same latent space, where the inner product between the user and item vector 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 [?] introduces real-valued features to MF so that the sparsity is alleviated. 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 relies on some manual features such as topics [?], geographics [?], and demographics [?] to model news and users. In the recent ten years, deep learning shows an unlimited potential and prevails in the task of representation learning [?]. 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. [? ? ? ? ? ? ?] proposes effective models that employs CNN, RNN, multi-head attention, and personalized attention to represent news and users. External information such as knowledge and user-item biparititie graph are also incorporated to enhance representations.

More recently, large-scale pretrained language models (PLM in short) e.g. BERT [?] 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 [?] optimizes the training scheme to improve model efficiency. The model deduplicates user's historical news and candidate news in a batch, significantly speeding up the training process. Another line of research modifies the self-attention layer to reduce complexity. For example, Longfomer []; Fastformer aggregates self-attended query into a global one and make the complexity linear.

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