A Appendix

A.1 Dataset

As mentioned in the main content, we evaluate our model performance on three public session search log datasets: AOL4PS log data [11], TREC Session Track 2014 log data [11, 12] and TianGong-ST query log data [17] query log data. The basic information of the three datasets is presented in Table 1. We here present the details of the datasets.

AOL4PS. The AOL4PS dataset [25] is collected and processed from AOL query logs [39]. AOL Search released a collection of user query logs that include a large number of queries from 657,426 users over a three-month period [39]. AOL query logs only contain the records of user-clicked documents for the queries and do not have the records of the candidate documents returned by the search engine. Thus, methods were presented to obtain the candidate results for the queries. The provider of AOL4PS optimized the process of data set construction and proposed an improved BM25 algorithm, which proved to be useful.

TREC 2014 Search Log. The TREC 2014 Session Track Search Log [11] is collected from a search system based on the indri index of ClueWeb12 by the track organizers. The session logs contain all URLs, snippets, clicks, and titles of search results to the user. Notably, for each entered query in sessions, the Search Log lists 10 candidate documents ranked by the search engine. We split the whole dataset into a train set, validation set, and test set by the ratio of 5:1:1. Each set contains several integral sessions. As suggested in [33], we use the snippets as the document of results.

TianGong-ST Search Log. TianGong-ST Search Log is extracted from a query log collected by Sogou.com[7], a Chinese commercial search engine. It presents the 18-day user-entered queries, the top-10 results returned from search engines, and the click interaction from the user. As it is claimed in [17], the dataset provider refine the sessions through steps including filtering sessions that contain pornographic, violent, or politically sensitive contents. We follow [14] to take the title of the search result as its content. Similar to our approach to the TREC dataset, we split the whole dataset into a train set, validation set, and test set by the ratio of 5:1:1.

A.2 Baselines

Our utilized baselines contain the popular choosed SNRM [57], HBA-Transformers [42], COCA [60], CARS [4], HEXA [52] and RICR [14]. We here present the details of our datasets.

A.2.1 Ad-hoc Methods. **SNRM** [57] is an ad-hoc method. It applies a standalone neural ranking model by introducing a sparsity property to learn a latent sparse representation for each query and document. This work uses a neural network based on n-gram representation learning which can capture the semantic relationship between the query and documents. As for the score function, they apply cosine similarity between the representations

A.2.2 Context-aware Methods. **HBA-Transformers** [42] and the below methods are context-aware. It is a popular choice for baseline due to its state-of-the-art performance. It introduces behavior awareness to a BERT-based ranker. The input of the model concatenates all search behaviors of the session into a sequence. Then, HBA-Transformers utilizes a hierarchical structure of an intra-behavior attention layer and an inter-behavior attention layer for better interaction modeling. The final score is calculated by the token [CLS] of the output representation and a linear classifier.

CARS [4] is a classic representation-based method in context-aware document ranking tasks. It introduces a two-level hierarchical recurrent neural network to learn the search context representation of individual queries, search tasks, and corresponding dependency structure. Beyond this, to identify variable dependency structure between search context and users' ongoing search activities, attention at both levels of recurrent states is leveraged.

COCA [60] utilizes contrastive learning to improve the sequence representation of BERT for document ranking. COCA aims to learn a more accurate representation of the user interaction sequence by considering the possible variations in user interactions. The COCA model was used in NUCIR-16 Session Search Task as a backbone and outperforms all other participants' runs in the task [13]. Notably, we only utilize the origin ranking method of COCA.

HEXA [52] exploits heterogeneous graphs to organize the contextual information and beneficial search logs for modeling user intents and ranking results. HEXA constructs a session graph built from the current session queries and documents and a query graph by sampling the current query's k-layer neighbors from search logs. The heterogeneous graph neural networks are utilized for enhancing the ranking process. HEXA is based on the interaction-based framework.

RICR [14] encodes the session history into a latent representation and uses this representation to enhance the current query and the candidate document. It then matches the enhanced query and candidate document with several matching components to capture the ingrained information of word-level interactions. This method shows comparable performance against the other context-aware methods. Notably, since each query may have multiple corresponding clicked documents in the TREC Session Track 14 dataset which makes RICR [14] unable to be implemented, thus we do not report its results on this dataset.

[7]https://www.sougou.com

References

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* 19–26.

[2] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. 2006. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.* 3–10.

[3] Wasi Uddin Ahmad and Kai-Wei Chang. 2018. Multi-task learning for document ranking and query suggestion. In *Sixth International Conference on Learning Representations.*

[4] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 385–394.

[5] Paul N Bennett, Krysta Svore, and Susan T Dumais. 2010. Classification-enhanced ranking. In *Proceedings of the 19th international conference on World wide web.* 111–120.

[6] Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* 185–194.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning.* 89–96.

[8] Fei Cai, Shangsong Liang, and Maarten De Rijke. 2014. Personalized document re-ranking based on bayesian probabilistic matrix factorization. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* 835–838.

[9] Huanhuan Cao, Daxin Jiang, Jian Pei, Enhong Chen, and Hang Li. 2009. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web.* 191–200.

[10] Mark J Carman, Fabio Crestani, Morgan Harvey, and Mark Baillie. 2010. Towards query log based personalization using topic models. In *Proceedings of the 19th ACM international conference on Information and knowledge management.* 1849–1852.

[11] Ben Carterette, Paul Clough, Mark Hall, Evangelos Kanoulas, and Mark Sanderson. 2016. Evaluating retrieval over sessions: The TREC session track 2011-2014. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* 685–688.

[12] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track.* Technical Report. DELAWARE UNIV NEWARK DEPT OF COMPUTER AND INFORMATION SCIENCES.

[13] Haonan Chen and Zhicheng Dou. 2022. RUCIR at the NTCIR-16 Session Search (SS) Task. *Proceedings of NTCIR-16* (2022).

[14] Haonan Chen, Zhicheng Dou, Qiannan Zhu, Xiaochen Zuo, and Ji-Rong Wen. 2023. Integrating Representation and Interaction for Context-Aware Document Ranking. *ACM Transactions on Information Systems* 41, 1 (2023), 1–23.

[15] Haonan Chen, Zhicheng Dou, Yutao Zhu, Zhao Cao, Xiaohua Cheng, and Ji-Rong Wen. 2022. Enhancing user behavior sequence modeling by generative tasks for session search. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management.* 180–190.

[16] Jia Chen, Jiaxin Mao, Yiqun Liu, Ziyi Ye, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. 2021. A Hybrid Framework for Session Context Modeling. *ACM Transactions on Information Systems (TOIS)* 39, 3 (2021), 1–35.

[17] Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. TianGong-ST: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management.* 2485–2488.

[18] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining.* 126–134.

[19] Michel Deudon. 2018. Learning semantic similarity in a continuous space. *Advances in neural information processing systems* 31 (2018).

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web.* 581–590.

[22] Georges Dupret and Ciya Liao. 2010. A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In *Proceedings of the third ACM international conference on Web search and data mining.* 181–190.

[23] Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. 2018. Personalizing search results using hierarchical RNN with query-aware attention. In *Proceedings of the 27th ACM international conference on information and knowledge management.* 347–356.

[24] Antoine Gourru, Julien Velcin, and Julien Jacques. 2021. Gaussian embedding of linked documents from a pretrained semantic space. In *Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).*

[25] Qian Guo, Wei Chen, and Huaiyu Wan. 2021. AOL4PS: A Large-scale Data Set for Personalized Search. *Data Intelligence* 3, 4 (10 2021), 548–567. https://doi.org/10.1162/dint_a_00104 arXiv:https://direct.mit.edu/dint/article-pdf/3/4/548/1968580/dint_a_00104.pdf

[26] Morgan Harvey, Fabio Crestani, and Mark J Carman. 2013. Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* 2309–2314.

[27] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* 2333–2338.

[28] Shan Jiang, Yuening Hu, Changsung Kang, Tim Daly Jr, Dawei Yin, Yi Chang, and Chengxiang Zhai. 2016. Learning query and document relevance from a web-scale click graph. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* 185–194.

[29] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* 133–142.

[30] Rosie Jones and Kristina Lisa Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management.* 699–708.

[31] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864* (2020).

[32] Jianghao Lin, Weiwen Liu, Xinyi Dai, Weinan Zhang, Shuai Li, Ruiming Tang, Xiuqiang He, Jianye Hao, and Yong Yu. 2021. A graph-enhanced click model for web search. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval.* 1259–1268.

[33] Aldo Lipani, Ben Carterette, and Emine Yilmaz. 2019. From a user model for query sessions to session rank biased precision (sRBP). In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval.* 109–116.

[34] Xiangyong Liu, Guojun Wang, and Md Zakirul Alam Bhuiyan. 2022. Personalised context-aware re-ranking in recommender system. *Connection Science* 34, 1 (2022), 319–338.

[35] Yiding Liu, Weixue Lu, Suqi Cheng, Daiting Shi, Shuaiqiang Wang, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained language model for web-scale retrieval in baidu search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.* 3365–3375.

[36] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[37] Shuqi Lu, Zhicheng Dou, Xu Jun, Jian-Yun Nie, and Ji-Rong Wen. 2019. Psgan: A minimax game for personalized search with limited and noisy click data. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval.* 555–564.

[38] Shengjie Ma, Chong Chen, Jiaxin Mao, Qi Tian, and Xuhui Jiang. 2023. Session Search with Pre-trained Graph Classification Model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 953–962.

[39] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st international conference on Scalable information systems.* 1–es.

[40] Daniel Poh, Bryan Lim, Stefan Zohren, and Stephen Roberts. 2022. Enhancing cross-sectional currency strategies by context-aware learning to rank with self-attention. *The Journal of Financial Data Science* (2022).

[41] Chen Qian, Fuli Feng, Lijie Wen, and Tat-Seng Chua. 2021. Conceptualized and contextualized gaussian embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13683–13691.

[42] Chen Qu, Chenyan Xiong, Yizhe Zhang, Corby Rosset, W Bruce Croft, and Paul Bennett. 2020. Contextual re-ranking with behavior aware transformers. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1589–1592.

[43] Xuehua Shen, Bin Tan, and ChengXiang Zhai. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* 43–50.

[44] Yang Song, Hongning Wang, and Xiaodong He. 2014. Adapting deep ranknet for personalized search. In *Proceedings of the 7th ACM international conference on Web search and data mining.* 83–92.

[45] David Sontag, Kevyn Collins-Thompson, Paul N Bennett, Ryen W White, Susan Dumais, and Bodo Billerbeck. 2012. Probabilistic models for personalizing web search. In *Proceedings of the fifth ACM international conference on Web search and data mining.* 433–442.

[46] Zhan Su, Zhicheng Dou, Yujia Zhou, Ziyuan Zhao, and Ji-Rong Wen. 2023. PSLOG: Pretraining with Search Logs for Document Ranking. (2023).

[47] Zhan Su, Zhicheng Dou, Yutao Zhu, Xubo Qin, and Ji-Rong Wen. 2021. Modeling intent graph for search result diversification. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 736–746.

[48] Jaime Teevan, Daniel J Liebling, and Gayathri Ravichandran Geetha. 2011. Understanding and predicting personal navigation. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 85–94.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[50] Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623* (2014).

[51] Thanh Vu, Dat Quoc Nguyen, Mark Johnson, Dawei Song, and Alistair Willis. 2017. Search personalization with embeddings. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*. Springer, 598–604.

[52] Shuting Wang, Zhicheng Dou, and Yutao Zhu. 2023. Heterogeneous Graph-based Context-aware Document Ranking. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 724–732.

[53] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd international conference on World Wide Web*. 1411–1420.

[54] Qiang Wu, Chris JC Burges, Krysta M Svore, and Jianfeng Gao. 2008. *Ranking, boosting, and model adaptation*. Technical Report. Citeseer.

[55] Yuhang Ye, Zhonghua Li, Zhicheng Dou, Yutao Zhu, Changwang Zhang, Shangquan Wu, and Zhao Cao. 2023. Learning from the Wisdom of Crowds: Exploiting Similar Sessions for Session Search. (2023).

[56] Arda Yüksel, Berke Uğurlu, and Aykut Koç. 2021. Semantic change detection with gaussian word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3349–3361.

[57] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik 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*. 497–506.

[58] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. 2005. Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 504–511.

[59] Cheng Zhang, Qiuchi Li, Lingyu Hua, and Dawei Song. 2020. Assessing the Memory Ability of Recurrent Neural Networks. *arXiv preprint arXiv:2002.07422* (2020).

[60] Yutao Zhu, Jian-Yun Nie, Zhicheng Dou, Zhengyi Ma, Xinyu Zhang, Pan Du, Xiaochen Zuo, and Hao Jiang. 2021. Contrastive learning of user behavior sequence for context-aware document ranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2780–2791.