GenSAR: Unifying Balanced Search and Recommendation with Generative Retrieval

Anonymous Author(s)

ACM Reference Format:
Anonymous Author(s). 2018. GenSAR: Unifying Balanced Search and Recommendation with Generative Retrieval. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

A Experimental Details

In this section, we introduce the experimental details.

A.1 Dataset

Since GenSAR requires both user S&R interaction logs, as well as textual information about items, no publicly available dataset, to the best of our knowledge, contains all these elements. While KuaiSAR [23] includes S&R logs, it lacks textual information. Consequently, we conducted experiments on the following datasets: a modified version of the Amazon dataset and a dataset collected from a Chinese commercial app.

Amazon[1] [9, 14]: We adopted the widely accepted semi-synthetic Amazon dataset. Following previous studies [2, 3, 20, 21], we generated synthetic search data based on the existing recommendation data. Since many items in the "Kindle Store" subset used in prior work [20, 21] lack textual information, we instead utilized the "Electronics" subset. We selected the five-core subset, where all users and items have at least five interactions. Following the data construction method used in previous studies [20, 21], we applied the leave-one-out strategy to split the dataset into training, validation, and test sets.

Commercial: To thoroughly evaluate the effectiveness of GenSAR, we collected a dataset from a Chinese commercial app, containing S&R interactions from 5,000 users over two weeks. We applied five-core filtering, removing users and items with fewer than five interactions. We also scraped textual information for the items, including titles and descriptions. Since the raw text contained noise, we utilized Qwen-2.5[2] [28] to summarize the text for each item, effectively filtering out irrelevant information. We sorted the data chronologically and split it into training, validation, and test sets in an 8:1:1 ratio.

[1] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html, https://github.com/QingyaoAi/Amazon-Product-Search-Datasets
[2] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

A.2 Baselines

In this section, we introduce the details of the baselines.

First, we compare with the following recommendation models: (1) **Sequential Recommendation**: **GRU4Rec** [10] utilizes GRUs to capture users' interaction histories; **SASRec** [12] is a model based on a unidirectional Transformer; **FMLP-Rec** [33] is a MLP-based model with learnable filters; **LRURec** [29] uses Linear Recurrent Units. (2) **Generative Recommendation**: **P5-CID** [8, 11] integrates collaborative signals using spectral clustering on item co-appearance graphs; **TIGER** [18] employs RQ-VAE to create codebook-based identifiers, encoding semantic information into code sequences; **LC-Rec** [32] leverages codebook-based identifiers and auxiliary alignment tasks to link generated code sequences with natural language.

Next, we compare with the following search models: (1) **Personalized Search**: **QEM** [2] focuses solely on the matching scores between items and queries; **TEM** [4] leverages a Transformer encoder to aggregate the user's interacted items along with the current query; **CoPPS** [5] employs contrastive learning methods. (2) **Dense Retrieval**: **E5**[3] [25] is a multilingual text embedding model, extending the English E5, trained through contrastive pre-training on large text pairs and fine-tuned on high-quality labeled data; **BGE**[4] [26] provides versatile embedding models built on a BERT [7]-like architecture, offering a balance between performance and efficiency, with support for easy fine-tuning. (3) **Generative Retrieval**: **DSI-QG** [35] enhances DSI [24], a Transformer-based retrieval model, by incorporating query generation to address data distribution mismatches between indexing and retrieval; **WebUltron** [34] enhances DSI by employing product quantization for generating semantic IDs and utilizing URLs to construct term-based IDs. **GenRet** [22] learns to tokenize documents into docids via a discrete auto-encoding approach.

Finally, we compare with the following joint S&R models: **JSR** [30] is a comprehensive framework that optimizes a combined loss function; **SESRec** [21] employs contrastive learning to develop disentangled search representations for recommendation; **UnifiedSSR** [27] simultaneously learns user behavior history for both S&R scenarios; **UniSAR** [20] is a unified S&R model that models user transition behaviors.

A.3 Evaluation Metrics

To evaluate the performance of recommendation and search, following previous works [20, 21, 33], we use ranking metrics including top-k *Hit Ratio* (HR) and top-k *Normalized Discounted Cumulative Gain* (NDCG). We report the results for k values of {1, 5, 10}, and since NDCG@1 is the same as HR@1, we do not report it. For recommendation, following commonly used settings [20, 21, 33],

[3] https://huggingface.co/intfloat/multilingual-e5-base
[4] https://huggingface.co/BAAI/bge-base-en-v1.5, https://huggingface.co/BAAI/bge-base-zh-v1.5

we pair the ground-truth item with 99 randomly sampled items that the user has not interacted with to form the candidate list. For search, since semantic relevance is crucial, randomly sampled negative samples are likely to be semantically irrelevant, making them overly simple and ineffective for distinguishing model performance. To address this, following previous work [1, 6], we use BM25 [19] to retrieve 99 negative samples to construct the candidate list. These samples are more challenging (hard negatives), providing a better evaluation of model effectiveness.

A.4 Implementation Details

For identifier learning, the semantic embedding is obtained using BGE [26], while the collaborative embedding is derived from the trained UniSAR [20] model. The number of codebooks for both shared and specific (L_m and L_n) is set to 2. Each codebook contains 256 code embeddings. The code embeddings in the shared codebooks have a dimension of 64, while those in the specific codebooks have a dimension of 32. α in Eq. (**??**) is set to 0.25. We train the RQ-VAE model used for obtaining identifiers in Section **??** for 500 epochs using the Adam [13] optimizer with a learning rate of 1e-3 and a batch size of 1024.

For the LLM backbone in all models, we use "t5-small"[5] [17] for the Amazon dataset and "Randeng-T5-77M-MultiTask-Chinese"[6] [31] for the Commercial dataset. Following previous works [22, 35], we adopt the Doc2query [15, 16] technique. The Doc2query model is "msmarco-t5-base-v1"[7] for Amazon and "msmarco-chinese-mt5-base-v1"[8] for Commercial. For the Amazon dataset, we generate 1 pseudo-query for each item, and for the Commercial dataset, we generate 10 pseudo-queries for each item. The Doc2query model is fine-tuned on search data before generating pseudo-queries. The maximum historical length for S&R is set to 5. We train the LLM for S&R tasks, as described in Section **??**, using the Adam optimizer with an initial learning rate of 1e-3 and a cosine learning rate schedule. For all generative methods, we set the beam size to 30. All the experiments are conducted on 8 NVIDIA Tesla v100 GPUs.

References

[1] Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2018. Multi-task learning for document ranking and query suggestion. In *International conference on learning representations*.

[2] Qingyao Ai, Daniel N Hill, SVN Vishwanathan, and W Bruce Croft. 2019. A zero attention model for personalized product search. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 379–388.

[3] Qingyao Ai, Yongfeng Zhang, Keping Bi, Xu Chen, and W Bruce Croft. 2017. Learning a hierarchical embedding model for personalized product search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 645–654.

[4] Keping Bi, Qingyao Ai, and W Bruce Croft. 2020. A transformer-based embedding model for personalized product search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1521–1524.

[5] Shitong Dai, Jiongnan Liu, Zhicheng Dou, Haonan Wang, Lin Liu, Bo Long, and Ji-Rong Wen. 2023. Contrastive Learning for User Sequence Representation in Personalized Product Search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*. ACM, 380–389.

[5] https://huggingface.co/google-t5/t5-small
[6] https://huggingface.co/IDEA-CCNL/Randeng-T5-77M-MultiTask-Chinese
[7] https://huggingface.co/doc2query/msmarco-t5-base-v1
[8] https://huggingface.co/doc2query/msmarco-chinese-mt5-base-v1

[6] Chenlong Deng, Yujia Zhou, and Zhicheng Dou. 2022. Improving personalized search with dual-feedback network. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 210–218.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

[8] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[10] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[11] Wenyue Hua, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2023. How to index item ids for recommendation foundation models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 195–204.

[12] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[15] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6, 2 (2019).

[16] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[17] 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.

[18] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 10299–10315.

[19] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[20] Teng Shi, Zihua Si, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Dewei Leng, Yanan Niu, and Yang Song. 2024. UniSAR: Modeling User Transition Behaviors between Search and Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1029–1039.

[21] Zihua Si, Zhongxiang Sun, Xiao Zhang, Jun Xu, Xiaoxue Zang, Yang Song, Kun Gai, and Ji-Rong Wen. 2023. When search meets recommendation: Learning disentangled search representation for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1313–1323.

[22] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2024. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2024).

[23] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Dewei Leng, Yanan Niu, Yang Song, Xiao Zhang, and Jun Xu. 2023. KuaiSAR: A Unified Search And Recommendation Dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 5407–5411.

[24] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.

[25] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672* (2024).

[26] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 641–649.

[27] Jiayi Xie, Shang Liu, Gao Cong, and Zhenzhong Chen. 2024. UnifiedSSR: A Unified Framework of Sequential Search and Recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3410–3419.
[28] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).
[29] Zhenrui Yue, Yueqi Wang, Zhankui He, Huimin Zeng, Julian McAuley, and Dong Wang. 2024. Linear recurrent units for sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 930–938.
[30] Hamed Zamani and W. Bruce Croft. 2018. Joint Modeling and Optimization of Search and Recommendation. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems, Bertinoro, Italy, August 28-31, 2018 (CEUR Workshop Proceedings, Vol. 2167)*. CEUR-WS.org, 36–41.
[31] Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR* abs/2209.02970 (2022).
[32] Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 1435–1448.
[33] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-Enhanced MLP is All You Need for Sequential Recommendation. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2388–2399.
[34] Yujia Zhou, Jing Yao, Ledell Wu, Zhicheng Dou, and Ji-Rong Wen. 2023. WebUltron: An Ultimate Retriever on Webpages Under the Model-Centric Paradigm. *IEEE Transactions on Knowledge and Data Engineering* (2023).
[35] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128* (2022).