Enhancing Multi-Domain Recommendations via LLM-Generated Data

Chumeng Jiang 2024316092 Kairong Luo 2024310643 Zhixuan Pan 2024311550

1 Introduction

Due to the increasingly severe information overload issue, the application of recommendation systems (RS) prevails across all kinds of Internet platforms, as they can provide personalized items for each user. Recently, with the growing number of specialized domains in one comprehensive platform, such as short video recommendations, article recommendations, and product recommendations in the same app, multi-domain recommendation has garnered significant attention. The multi-domain recommendation can simultaneously leverage knowledge from different domains, alleviating the data sparsity issue and allowing a single model to make recommendations across multiple domains, reducing the deploying costs.

Nevertheless, the historical data size between different domains varies. Some areas may have significantly more data than others, namely the rich or cold-start scenarios. This disparity in data size can lead to certain limitations during model training. For instance, the learning of domain-specific parameters in cold-start scenarios may be insufficient, while the learning of domain-shared parameters may be dominated by rich scenarios. Previous work mainly addresses these issues through meticulous structural design of the models [5, 1, 13, 17].

In this paper, we adopt a different perspective by addressing this issue from the data standpoint. The emergence of LLMs has made it possible to generate virtual user and item data. Furthermore, LLMs, with their extensive world knowledge and outstanding comprehending capability, have demonstrated impressive recommendation capabilities in cold-start scenarios [3]. In this case, we utilize the LLMs to simulate users in cold-start scenarios and synthesize more sufficient positive samples after learning from existing multi-domain historical interactions. Through an elaborately designed data filtering and denoising strategy, the recommendation quality of the multi-domain models can be enhanced. Moreover, through the lens of recommendation systems, we may get more insight into the synthetic data from existing LLMs [12, 10].

2 Related Work

2.1 Multi-Domain Recommendation

Extensive work has been done to address the challenge of recommendation within multi-domain scenarios. The multi-domain recommendation is a cross-domain recommendation approach aimed at enhancing accuracy across multiple domains simultaneously. For instance, Domain generalization (DG) methods extract common knowledge from various domains, thereby improving generalization to unknown domains and mitigating data sparsity issues [17]. STAR [13] enhances the capture of the features of each domain by introducing Partitioned Normalization (PN) and domain-specific fully connected networks (FCNs) to capture the unique characteristics of each domain.

However, many of these methods struggle with effectively disentangling domain-shared and domain-specific knowledge. Methods like CATART [5] utilize auto-encoders to create global embeddings from domain-specific ones, relying on attention mechanisms to aggregate these embeddings for recommendations. However, this approach can inadvertently compromise disentanglement, as domain-shared knowledge influences the updating of domain-specific embeddings. Similarly, SAML [1]

Preprint. Under review.

maps features into global and domain-specific embeddings and employs a mutual unit to learn domain similarity, yet it lacks effective alignment mechanisms and fails to fully exploit the inter-domain relationships. Overall, while various methodologies exist, challenges related to model complexity and data sparsity persist.

2.2 Data Synthesis in Recommendation

Many traditional synthetic data generation methods have been proposed to address data imbalance in RS, i.e. the issues of the data sparsity and the long-tail distribution in the data. Common methods [14] include using k-nearest neighbors to create new instances based on existing minority class samples, employing generative models like GANs, VAEs, and diffusion models to generate synthetic tabular data. The data synthesized by these traditional methods often closely resembles the distribution of the original dataset.

Using LLMs to synthesize data allows for the incorporation of LLMs' inherent world knowledge and the utilization of more textual information during the synthesis process. Synthetic data generated by LLMs has been applied in fields such as healthcare [7, 16], demonstrating certain effectiveness. In the recommendation field, attempts to utilize data synthesized by LLMs are still relatively limited. ONCE [9] prompts closed-source LLMs to synthesize the items that new users with fewer than five historical records might be interested in, enriching their data to achieve better user embeddings. LLMRec [15] conducts the data augmentation by requiring the LLMs to select the most likely positive and negative item pairs within the candidate set provided. It leverages LLMs' advantages in pointwise comparison and employs MAE and noise pruning for denoising. However, there is no previous work that utilizes LLM-generated data to promote multi-domain recommendation. There are challenges including balancing the data density across multiple domains, controlling the noises introduced from synthetic data, and aligning the LLMs with the multi-modal data distributions in the multi-domain RS scenarios.

3 Problem Formulation

We target one of the most common problems in multi-domain recommendation: CTR prediction. The recommender system uses interaction data $(d, \mathbf{x})$ to predict y under the multi-domain setting, where y indicates whether a click occurred ($y = 1$) or not ($y = 0$). Here, d is the domain indicator, distinguishing samples from D total domains, and $\mathbf{x}$ consists of raw features, including user and item attributes. Assuming there are M categorical features, $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]$, with $\mathbf{x}_m$ being the one-hot representation of the m-th feature. An embedding layer maps $\mathbf{x}$ to a low-dimensional vector $\mathbf{e} = [\mathbf{e}_1 \parallel \mathbf{e}_2 \parallel \ldots \parallel \mathbf{e}_M]$, where $\parallel$ indicates concatenation. For the m-th feature, $\mathbf{e}_m$ is derived through a learnable lookup operation $\mathbf{e}_m = \mathbf{E}_m \cdot \mathbf{x}_m$, with $\mathbf{E}_m \in \mathbb{R}^{k \times \mu_m}$ as the weight matrix, k as the embedding size and u_m as the number of feature values. Finally, the predicted result $\hat{y}$ for whether a user will click on an item is computed as $\hat{y} = f_d(\mathbf{e})$, where f_d is the recommendation model for the d-th domain. We use the Negative Log-Likelihood Loss, which is also known as the binary cross-entropy loss $\mathcal{L}(\boldsymbol{\Theta}) = -\frac{1}{B}\sum_{i=1}^{B}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$, where Θ represents the learnable parameters, B is the batch size, and y_i and $\hat{y}_i$ are the true label and predicted result for the i-th sample, respectively.

4 Method

- **Datasets:** The *Amazon Datasets*[1] consist of product and user metadata, and users' product reviews from the Amazon platform. There are 29 categories in total, with some overlapping users across different categories. Recommendations for products in each distinct category can be viewed as a separate domain.
- **Recommendation Models:** (1) Single-domain models: SASRec [4] and LightGCN [2]; (2) Multi-domain models: EMCDR [11], BiTCDR [8] and CUT [6].
- **Evaluation Metrics:** We choose *AUC* as the measure of recommendation accuracy, while implementing the *Gini coefficient* and *item coverage* to examine the user-side and item-side recommendation fairness, respectively.

[1] https://nijianmo.github.io/amazon/index.html

References

[1] Yuting Chen, Yanshi Wang, Yabo Ni, An-Xiang Zeng, and Lanfen Lin. Scenario-aware and mutual-based approach for multi-scenario recommendation in e-commerce, 2020.

[2] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation, 2020.

[3] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *ECIR*, 2024.

[4] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018.

[5] Chenglin Li, Yuanzhen Xie, Chenyun Yu, Bo Hu, Zang Li, Guoqiang Shu, Xiaohu Qie, and Di Niu. One for all, all for one: Learning and transferring user embeddings for cross-domain recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 366–374, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Hanyu Li, Weizhi Ma, Peijie Sun, Jiayu Li, Cunxiang Yin, Yancheng He, Guoqiang Xu, Min Zhang, and Shaoping Ma. Aiming at the target: Filter collaborative information for cross-domain recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2081–2090, New York, NY, USA, 2024. Association for Computing Machinery.

[7] Che Liu, Zhongwei Wan, Haozhe Wang, Yinda Chen, Talha Qaiser, Chen Jin, Fariba Yousefi, Nikolay Burlutskiy, and Rossella Arcucci. Can medical vision-language pre-training succeed with purely synthetic data?, 2024.

[8] Meng Liu, Jianjun Li, Guohui Li, and Peng Pan. Cross domain recommendation via bi-directional transfer graph collaborative filtering networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 885–894, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 452–461, 2024.

[10] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data for language models. *arXiv preprint arXiv:2404.07503*, 2024.

[11] Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. Cross-domain recommendation: an embedding and mapping approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2464–2470. AAAI Press, 2017.

[12] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

[13] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, and Xiaoqiang Zhu. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4104–4113, New York, NY, USA, 2021. Association for Computing Machinery.

[14] Fatih Cihan Taskin, Ilknur Akcay, Muhammed Pesen, Said Aldemir, Ipek Iraz Esin, and Furkan Durmus. Effects of using synthetic data on deep recommender models' performance. *arXiv preprint arXiv:2406.18286*, 2024.

[15] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 806–815, 2024.

[16] Zerui Xu, Fang Wu, Tianfan Fu, and Yue Zhao. Retrieval-reasoning large language model-based synthetic clinical trial generation, 2024.

[17] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4396–4415, April 2023.