Semantic-Aware Data Augmentation for Sequential Recommendation

Zhifu Wei

Northeastern University, China
20235980@stu.neu.edu.cn

Abstract

Sequential recommendation (SR) aims to model users' dynamic preferences from their historical interaction sequences to provide personalized recommendations. However, data sparsity remains a core bottleneck limiting the performance of sequential recommendation models. Existing mixup methods face two major challenges: 1) They cannot effectively address the data sparsity dilemma in long-tail scenarios. 2) It is difficult to maintain the Semantic structure of augmented samples during the random mixing process. To address these challenges, this study proposes the **S**emantic-**A**ware **D**ata **A**ugmentation (SADA) framework, which utilizes large language models (LLMs) to generate semantic embeddings. This framework allows for the fusion of both collaborative and semantic signals, alleviating the representation deficiency of long-tail items. Additionally, through semantic-guided mixup, the framework preserves semantic structure consistency at both the user and item levels, thereby avoiding semantic structure degradation caused by traditional random mixing. This approach is expected to significantly improve recommendation performance and generalization ability across multiple datasets and application scenarios. In a broader context, this research aims to drive the evolution of data augmentation in sequential recommendation from heuristic methods to a semantic-driven paradigm, helping to build more personalized, accurate, and socially valuable recommendation services.

Introduction

In the era of information overload, sequential recommendation has become a key bridge connecting users with vast amounts of content (Kang and McAuley 2018; Meng et al. 2023). By accurately capturing user preferences and providing personalized recommendations, these systems have created significant commercial value in various fields, such as e-commerce, streaming, and social networks. However, data sparsity in sequential recommendation has become increasingly prominent. The explosive growth of platform data, limitations in user interactions, and cross-platform data silos and privacy constraints have collectively led to a scarcity of usable data (Jing et al. 2023; Chen et al. 2024).

To address these challenges, many recent studies have started exploring the potential of mixup techniques in sequential recommendation. Mixup generates new training samples by mixing every interaction position of two user sequences at the representation level, thereby alleviating the data sparsity issue. (Yang et al. 2023). However, this method faces two key challenges: 1) **Data scarcity dilemma:** In mixup, linear interpolation is required based on the existing data, but for long-tail items with extremely sparse interactions, this demand creates a "data-dependent data augmentation" cycle dilemma. 2) **Semantic structure disruption:** Traditional representation mixing methods often rely on random strategies, such as randomly selecting different sequences within a batch for mixing or randomly sampling mixing weights from a beta distribution. However, different user sequences may exhibit entirely distinct interest biases, and the attributes of the items they interact with could vary significantly as well. This randomness often leads to severe disruption of the sequence's semantic structure.

Recent developments in large language models (LLMs) have made a breakthrough in alleviating data sparsity. LLMs can generate high-quality item semantic embeddings based on user and item attribute inputs (Liu et al. 2025). Building on this, we propose the **S**emantic-**A**ware **D**ata **A**ugmentation (SADA) framework, which incorporates collaborative-semantic dual signal fusion to address the sparse collaborative signals of long-tail items. To prevent the disruption of semantic structures, we leverage user sequence semantic representations generated by LLMs for sequence clustering, ensuring mixup occurs only within user clusters with similar interests. In addition, we adaptively adjust the mixing weights between items based on their semantic relevance, ensuring that the fusion of dissimilar items is minimized by reducing the mixing intensity.

Background

LLMs for Sequential Recommendation

In recent years, as large language models have developed, several works have emerged exploring how to leverage LLMs in sequential recommendation. LLM-ESR (Liu et al. 2025) improves the representation of long-tail users and items through dual-view modeling and self-distillation retrieval enhancement. LLM-Emb (Liu et al. 2024) proposes enhancing the embedding quality of extremely sparse items through supervised contrastive fine-tuning.

Mixup for Sequential Recommendation

With the widespread use of data augmentation in sequential recommendation, several studies have demonstrated the significant impact of mixup. BASRec (Dang et al. 2025) proposes single-sequence and cross-sequence mixing augmentation at the representation level. IDEA (Liao et al. 2025) generates multi-environment augmented samples by linearly mixing subsequences at the representation level. However, these methods still rely on random mixing strategies and do not fully account for user interest structures and item semantic associations, leading to the introduction of semantic noise and structural disruption during augmentation.

Approach

We propose the **SADA**, which consists of three modules designed to integrate collaborative and semantic signals, cluster user-interest structures, and apply adaptive mixing strategies. The specific methods are as follows:

Collaborative-Semantic Dual-Signal Fusion

To improve item representation quality when interaction data is sparse, we introduce semantic embeddings as an important complement to collaborative embeddings . Specifically, we use a large model API to generate semantic embeddings based on the textual attributes of items, which are then fused with the original ID embeddings before being input to the encoder (Liu et al. 2024; Hu et al. 2025). Given the significant differences in interaction frequencies between popular and long-tail items, and the fact that learning collaborative embeddings for long-tail items is often insufficient, we assign higher-weight semantic embeddings to long-tail items to compensate for the lack of collaborative signals. By combining collaborative and semantic dual signals, item representations can be effectively enhanced even when system behavior signals are extremely sparse.

User Sequence Clustering

To avoid randomly mixing user sequences with drastically different interests in mixup(Liao et al. 2025), we input user interaction sequences into an LLM, then perform K-means clustering on the generated user sequence semantic embeddings. Mixup is performed only within clusters or adjacent clusters of users with similar interest structures, ensuring consistency in the augmented samples' interest levels and preventing semantic drift caused by random strategies.

Adaptive Mixing Strategy

Additionally, to avoid forcibly mixing item representations with significant semantic differences during item-level mixing (Dang et al. 2025), which may disrupt the accurate semantic structure of the sequence, we leverage valuable prior knowledge from the large model to construct semantic similarity between items based on their attributes. Before performing item-level mixup, we adaptively set the mixing weights based on the semantic relevance between items: the lower the relevance, the weaker the mixing strength, ensuring that the generated item representations effectively maintain consistency with the original semantics.

Evaluation

We will evaluate the proposed sequence recommendation data augmentation method on public datasets and compare it with three categories of baselines: (1) **Backbone models:** GRU4Rec (Hidasi et al. 2015), SASRec (Kang and McAuley 2018), and FMLPRec (Zhou et al. 2022); (2) **Mixup-Based models:** BASRec (Dang et al. 2025) and IDEA (Liao et al. 2025). (3) **LLM-Based models:** LLM-ESR (Liu et al. 2025), LLM-Emb (Liu et al. 2024). Evaluation metrics mainly include HR@K and NDCG@K, with datasets partitioned using the leave-one-out strategy.

Experiments encompassed main comparative tests to evaluate overall performance against baselines across different datasets, along with ablation studies and similarity analyses to examine module contributions and the quality of augmented data. The success criteria were defined as achieving significant improvements over baseline methods in HR@K and NDCG@K across multiple datasets, while generating augmented data that maintains both relevance and diversity.

Discussion

Academic Value. Compared with the existing methods, our SADA has the following advantages: (i) **Collaborative–Semantic Dual Fusion:** SADA uses high-quality item semantic embeddings generated by large models as an effective supplement to collaborative information, overcoming the limitations of traditional augmentation methods, which are often constrained by insufficient collaborative signals in extremely sparse scenarios; (ii) **Sub-sequence Level Semantic Preservation Augmentation:** SADA ensures the overall coherence of sequence semantics during augmentation by performing sequence semantic clustering at the user level and applying semantic relevance weighting at the item level, thereby maintaining the integrity of sequence semantics throughout the mixup process. **Social Value.** (i) Significantly enhances recommendation quality and personalization on e-commerce, media, and social platforms, alleviating information overload; (ii) Improves exposure for long-tail items, thereby mitigating "filter bubbles"; (iii) Reduces data collection costs and enhances recommendation system universality through intelligent sample augmentation.

Conclusion

This study proposes a semantic-aware data augmentation framework, **SADA**, for sequential recommendation. The framework effectively alleviates the poor embedding quality of long-tail items by integrating collaborative signals with semantic embeddings generated by large models. Additionally, it introduces semantic-guided mixup at both the user sequence and item representation levels, ensuring that the augmentation process better preserves interest structure and semantic consistency, thereby generating more realistic sequence samples. This approach is expected to significantly improve recommendation quality and generalization, offering new insights to enhance the robustness and social value of sequential recommendations in real-world scenarios.

References

Chen, Y.; Zhao, S.; Wang, X.; Jin, Z.; Lin, Z.; and Wang, B. 2024. Towards robust data augmentation for sequence recommendation. In *CyberC*, 60–63.

Dang, Y.; Zhang, J.; Liu, Y.; Yang, E.; Liang, Y.; Guo, G.; Zhao, J.; and Wang, X. 2025. Augmenting Sequential Recommendation with Balanced Relevance and Diversity. In *AAAI*, volume 39, 11563–11571.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hu, G.; Zhang, A.; Liu, S.; Cai, Z.; Yang, X.; and Wang, X. 2025. AlphaFuse: Learn ID Embeddings for Sequential Recommendation in Null Space of Language Embeddings. In *SIGIR*, 1614–1623.

Jing, M.; Zhu, Y.; Zang, T.; and Wang, K. 2023. Contrastive self-supervised learning in recommender systems: A survey. *TOIS*, 42(2): 1–39.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *ICMD*, 197–206. IEEE.

Liao, Y.; Yang, Y.; Hou, M.; Wu, L.; Xu, H.; and Liu, H. 2025. Mitigating Distribution Shifts in Sequential Recommendation: An Invariance Perspective. In *SIGIR*, 1603–1613.

Liu, Q.; Wu, X.; Wang, W.; Wang, Y.; Zhu, Y.; Zhao, X.; Tian, F.; and Zheng, Y. 2024. Large Language Model Empowered Embedding Generator for Sequential Recommendation. *arXiv preprint arXiv:2409.19925*.

Liu, Q.; Wu, X.; Wang, Y.; Zhang, Z.; Tian, F.; Zheng, Y.; and Zhao, X. 2025. LLM-ESR: large language models enhancement for long-tailed sequential recommendation. In *NeurIPS*.

Meng, C.; Zhai, C.; Yang, Y.; Zhang, H.; and Li, X. 2023. Parallel knowledge enhancement based framework for multi-behavior recommendation. In *CIKM*, 1797–1806.

Yang, H.; Choi, Y.; Kim, G.; and Lee, J.-H. 2023. LOAM: improving long-tail session-based recommendation via niche walk augmentation and tail session mixup. In *SIGIR*, 527–536.

Zhou, K.; Yu, H.; Zhao, W. X.; and Wen, J.-R. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *WWW*, 2388–2399.