Transactions on Computational and Scientific Methods | Vo. 3, No. 2, 2023
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Multi-Level Attention and Sequence Modeling for Dynamic User Interest Representation in Real-Time Advertising Recommendation

Ming Wang
Northeastern University, San Jose, USA
mingwangcr@gmail.com

Abstract: This study investigates sequence modeling and attention optimization in real-time advertising recommendation, aiming to address the limitations of traditional methods in capturing dynamic user interests. It first analyzes the temporal dependencies of user behavior in advertising scenarios and points out that models based only on static correlations cannot fully represent the dynamic changes of user preferences. To solve this, a framework integrating sequence modeling with multi-level attention is proposed, which models user actions such as clicks, browsing, and purchases in order and achieves a unified representation of short-term interests and long-term preferences. The design introduces embedding layers to enhance feature representation and applies self-attention to highlight key behavior fragments, balancing global and local feature modeling. Residual connections and regularization are further incorporated to improve stability and generalization. Experiments compare the proposed method with multiple baselines on Precision@10, Recall@10, NDCG, and AUC, while also conducting sensitivity tests on regularization strength, model update intervals, cold-start ratios, and label noise. Results demonstrate that the proposed framework achieves consistent advantages across metrics, effectively capturing complex patterns in user behavior sequences and showing strong effectiveness and adaptability in real-time advertising recommendations.

Keywords: Real-time advertising recommendation; sequence modeling; attention mechanism; sensitivity analysis

1. Introduction

In the rapid development of the digital economy, advertising recommendations have become a core tool for internet companies to capture user attention and achieve business value. With the problem of information overload becoming increasingly severe, how to achieve personalized and precise recommendations in complex environments has become a central concern in both academia and industry. Recommendation systems no longer rely only on static features and simple correlation modeling[1]. They must be able to capture the dynamic patterns hidden in user behavior sequences to better understand real preferences. In this context, sequence modeling has gradually become a key direction in advertising recommendation research. Its goal is to exploit the temporal dependencies of user historical behaviors to improve the accuracy and robustness of recommendations.

Advertising recommendation scenarios are highly dynamic and complex. User actions such as clicks, browsing, searches, and purchases often exhibit strong temporal characteristics. These actions involve not only short-term local dependencies but also long-term contextual relationships[2]. For example, short-term clicks on a category of ads may relate to current interest trends, while long-term browsing records reveal more stable patterns of preference. Therefore, how to capture such cross-scale temporal dependencies effectively has become central to improving recommendation quality. Traditional modeling methods struggle with long-range dependencies and fail to fully exploit global information in historical behaviors. As a result, recommendation results often suffer from bias and local optima[3].

The introduction of attention mechanisms has provided new opportunities for sequence modeling in advertising recommendations. Attention can automatically focus on the most relevant parts of sequence data for the current task, avoiding the problem of information decay over long sequences. For advertising recommendations, attention not only assigns higher weight to key fragments in behavior sequences but also considers multi-level contextual relationships during modeling. This capability allows the system to extract core signals from complex behavior patterns, leading to more precise modeling of user preferences and significantly enhancing the level of personalization.

At the same time, advertising recommendation faces many challenges, such as sparse user behaviors, diverse interests, and rapidly changing external environments. A single modeling strategy is often insufficient in these conditions. It is therefore necessary to further optimize the combination of sequence modeling and attention mechanisms. Multi-level attention across different time scales and feature dimensions can strengthen adaptability to dynamic environments. Moreover, integrating structural optimization and regularization can reduce overfitting and noise interference, ensuring the stability and reliability of recommendation systems in practice[4].

Overall, sequence modeling and attention optimization in real-time advertising recommendations represent an important shift in research. The focus is moving from correlation-based modeling to causal and dynamic modeling. This shift improves both the accuracy and diversity of recommendations while also helping enterprises gain long-term user engagement and business value in a competitive digital ecosystem. From a theoretical perspective, such studies enrich the applications of sequence modeling and attention, driving a paradigm shift in recommendation system research. From a practical perspective, they provide more precise decision support for advertising strategies, promoting both user experience and commercial benefits. Thus, exploring how to optimize sequence modeling and attention mechanisms in real-time advertising recommendations has both high academic value and significant practical importance.

2. Related work

In recent years, advertising recommendation systems have shown a shift from traditional static modeling to dynamic modeling. Early methods were mainly based on collaborative filtering and matrix factorization. These approaches relied on user-item interaction matrices to learn latent features and could describe user preferences to some extent. However, they largely ignored the temporal dependencies of user behavior and relied only on static correlation modeling[5]. This limitation made it difficult to capture the dynamic changes of user interests over time in complex advertising scenarios. With the growth of data scale and computational power, sequence modeling has been introduced into advertising recommendation research to address these shortcomings in capturing behavioral dynamics.

In the field of sequence modeling, recurrent neural networks and their improved variants once played a dominant role. By modeling user behavior sequences step by step, these methods could capture short-term and medium-term dependencies to a certain degree and provide reasonable sequence representations for advertising recommendations. Yet recurrent structures face challenges in handling long sequences. Problems such as gradient vanishing and low computational efficiency limit their ability to balance global dependencies with local features. These issues have driven research toward more efficient deep modeling approaches, with flexible structures to capture long-term interest patterns and global preference characteristics[6].

With the development of deep learning, attention mechanisms have become central tools for overcoming bottlenecks in sequence modeling for advertising recommendations. The key advantage lies in their ability to assign weights automatically and highlight critical information in user behavior sequences. This avoids the decay of information in long sequences. Attention-based models can consider behavioral features across different time scales and strengthen interest representation through multi-dimensional feature interactions. In real-time advertising recommendation, the efficiency and interpretability of attention provide strong support for recommendation results and enable systems to adapt more effectively to rapid changes in user interests[7].

On this basis, the integration of sequence modeling and attention mechanisms has further advanced the optimization of advertising recommendations. Current research focuses not only on improving sequence modeling frameworks but also on introducing multi-level, cross-modal, and contextual strategies into attention mechanisms. These advances enhance the capacity of models to represent complex behavior patterns. At the same time, more work is being devoted to improving modeling efficiency in large-scale data environments, balancing real-time performance with accuracy, and enabling deployment in real advertising systems. Together, these developments are driving advertising recommendation from static prediction to dynamic, precise, and efficient intelligent recommendation, while also laying a solid foundation for future research.

3. Proposed Approach

In this study, the core idea of the method is to achieve efficient representation and modeling of user behavior sequences through sequence modeling and optimization of the attention mechanism. The overall model architecture is shown in Figure 1.

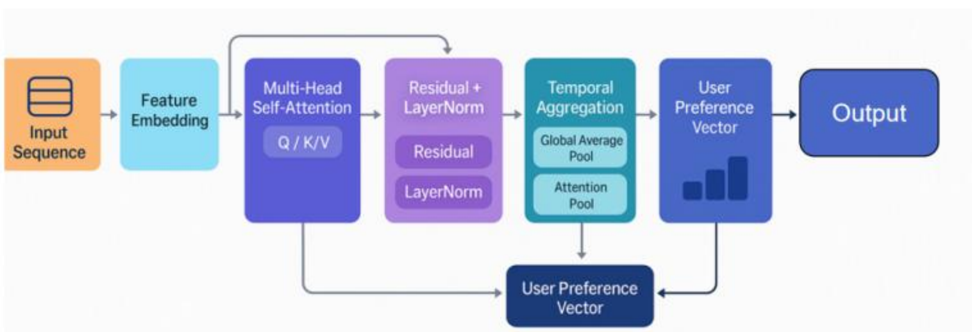

Figure 1. Overall model architecture

First, the user's historical behavior in the ad recommendation scenario is represented as an ordered sequence $X = \{x_1, x_2, ..., x_T\}$, where x_t represents the user's behavioral feature vector at time step t. To capture the potential temporal dependencies in the sequence, an embedding layer is constructed to map the original features into a high-dimensional space. Specifically, it is expressed as:

$$h_t = W_e \cdot x_t + b_e$$

Among them, W_e and b_e are the learnable weight matrix and bias term, respectively, and the obtained h_t represents the potential representation of the user at this time step.

On this basis, the self-attention mechanism is introduced to address the limitations of long sequence dependency modeling. For any sequence position t, the attention weight is calculated by mapping query, key, and value:

$$a_{ij} = \frac{\exp(\frac{q_i \cdot k^T}{\sqrt{d_k}})}{\sum_{j=1}^{T} \exp(\frac{q_i \cdot k^T}{\sqrt{d_k}})}$$

Among them, $q_i = h_i W_Q, k_j = h_j W_K, v_j = h_j W_V$, and the context representation of the sequence is obtained by weighted summation:

$$z_i = \sum_{j=1}^{T} a_{ij} v_j$$

This approach enables the model to adaptively distribute attention weights between different time steps, thereby focusing on behavioral segments that are more relevant to the recommendation task.

To better integrate global dependencies and local features, this method introduces residual connections and feedforward networks in the sequence modeling stage. For the context representation z_i at each position, the update rule is:

$$z_i = LayerNorm(z_i + h_i)$$

This is followed by a nonlinear transformation through a feedforward network:

$$o_i = RELU(\tilde{z}_i W_1 + b_1) W_2 + b_2$$

LayerNorm is used to stabilize the training process, while the feedforward network enhances the nonlinearity of feature representation. This design allows the model to simultaneously capture local details and global dependencies, thereby enhancing the richness of user interest representation.

Finally, to transform the sequence features into recommendation results, this study uses an aggregation mechanism to integrate the outputs of all time steps to obtain the user's final preference vector:

$$u = \frac{1}{T} \sum_{i=1}^{T} o_i$$

And through an output layer, the user interest vector is mapped to the candidate advertisement set to obtain the recommendation probability distribution:

$$\hat{y} = \mathrm{softmax}(uW_o + b_o)$$

Here, $\hat{y}$ represents the recommendation system's predicted probability for each ad. Through this modeling process, users' dynamic interests can be fully expressed, resulting in more accurate and adaptable recommendation results.

4. Experimental analysis

4.1 Dataset

This study uses the publicly available Taobao User Behavior (Tianchi) dataset as the data source. The dataset comes from real user behavior logs on a large e-commerce platform. It covers massive interactions between users and items and contains temporal records across multiple stages of user behavior. It is well-suited for building sequence modeling tasks in real-time advertising and recommendation scenarios. The core file is UserBehavior.csv, with fields including user_id, item_id, category_id, behavior_type (with values pv, fav, cart, buy), and a timestamp at the second level. The dataset is large in scale, fine in temporal granularity, and complete in behavior types. It reflects the multi-stage conversion path from exposure to click, favorite, add-to-cart, and purchase. It also naturally exhibits industrial characteristics such as sparsity and long-tail distribution.

Consistent with the objectives of real-time advertising recommendation, the dataset provides sufficient contextual information for sequential modeling. Multiple behavior sequences of the same user can be arranged chronologically. These sequences contain both short-term fluctuations of interest and long-term stable preferences. Based on the timestamp, features such as time periods, weekly cycles, and session windows can be derived. Based on category_id and historical interactions, features such as user-category preference and cross-behavior conversion rates can be constructed. The co-occurrence and ordering of multi-behavior signals in the temporal dimension provide reliable supervision and alignment anchors. These properties support attention mechanisms in focusing on key fragments and fusing information across different scales. The dataset, therefore, serves as a strong foundation for methods combining real-time modeling, sequential modeling, and attention optimization.

In terms of data usage, the raw logs are usually deduplicated and filtered for anomalies. They are then grouped by user and ordered by time to form session-based sequences. Training, validation, and test splits are created according to time to simulate online real-time scenarios. High-cardinality fields such as user_id, item_id, and category_id are processed by indexing or embedding. With multi-behavior labels, the dataset can support multi-task supervision, such as click prediction and conversion modeling, or cascaded tasks, such as recall, ranking, and reranking. Frequency thresholds can also be applied to alleviate cold-start and extreme sparsity issues. Overall, the Taobao User Behavior dataset combines openness, scale, and temporal completeness. It is an ideal benchmark for validating sequence modeling and attention optimization in real-time advertising recommendation.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1: Comparative experimental results

Method	Precision@10	NDCG	Recall@10	AUC
Sslrec[8]	0.412	0.398	0.365	0.721
RAKCR[9]	0.456	0.441	0.392	0.743
Mamba4rec[10]	0.489	0.472	0.418	0.768
HRLLM[11]	0.502	0.486	0.437	0.781
Ours	0.547	0.528	0.473	0.812

From the table results, the overall trend shows a gradual improvement. This indicates that with the advances in capturing sequential dependencies and optimizing attention mechanisms, both the accuracy and stability of advertising recommendations have been significantly enhanced. The improvements in Precision@10 and Recall@10 demonstrate that the model can achieve better alignment between user interests and candidate ads in short-list recommendations, which ensures more practical and reliable results. This progress not only reduces the disturbance of irrelevant ads but also confirms the necessity of introducing sequential modeling and attention mechanisms.

Furthermore, the increase in NDCG reflects the growing rationality of position ranking in the recommendation results. Compared with traditional methods, the improved model can better identify and emphasize key behavioral features, allowing highly relevant ads to appear earlier in the recommendation list. This optimization is especially important for advertising recommendations, since the system must not only provide relevant results but also ensure that high-value information is prioritized. This contributes to higher click-through rates and improved user experience. It shows that the proposed method has clear advantages in optimizing both recommendation quality and ranking structure.

The improvement in the AUC metric is also of great importance. It indicates that the model has enhanced overall discriminative ability. A higher AUC means the model is more robust and generalizable in distinguishing between ads that users are interested in and those they are not. This ability is critical in real-time advertising recommendation, where the system must adapt quickly to dynamic environments and diverse user groups. The improvement in AUC shows that the optimization can handle complex interactions between different users and ad contents, which strengthens adaptability in real-world applications.

Taken together, the performance across the four metrics shows that the proposed method achieves the best results in all dimensions. This demonstrates that the combination of sequence modeling and attention optimization can comprehensively capture dynamic user interests and cross-temporal dependencies. Such improvements represent not only a theoretical breakthrough but also a practical pathway for deployment in advertising recommendation systems. By effectively integrating sequential representations with attention mechanisms, the model can better capture user behaviors in real-time scenarios. This enhances both accuracy and diversity of recommendations and provides strong support for improving advertising effectiveness and user experience.

This paper also gives the influence of the regularization coefficient on the experimental results, and the experimental results are shown in Figure 2.

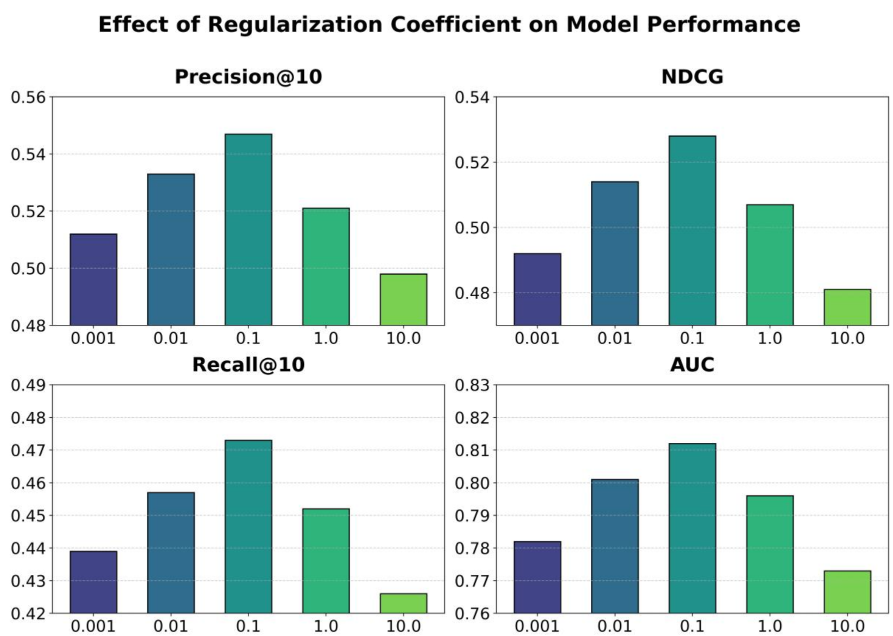

Figure 2. The influence of the regularization coefficient on experimental results

The results show that changes in the regularization coefficient have a significant impact on model performance. For Precision@10 and NDCG, when the coefficient increases from 0.001 to 0.1, the performance gradually improves and reaches the best value at 0.1. This indicates that moderate regularization can effectively suppress overfitting. It also improves both accuracy and ranking quality in short-list recommendation, enabling a more precise match between user interests and recommended ads.

A similar trend can be observed in Recall@10. When the regularization coefficient is too small, the model cannot constrain parameters effectively, leading to low recall. As the strength of regularization increases, the recall reaches its best value at 0.1, showing that the model can better balance information retention and redundancy suppression. However, when the coefficient continues to grow, recall drops sharply. This suggests that excessive constraints weaken the expressive power of user behavior features, preventing the system from capturing potential preferences comprehensively.

The performance of AUC further confirms this pattern. With an increasing regularization coefficient, the model's ability to discriminate between positive and negative samples gradually improves and peaks at 0.1. This reflects strong robustness and generalization ability. Yet, when the coefficient becomes too large, the discriminative power decreases. Overly strong regularization not only reduces the ability to suppress noise but also harms the representation of core features, which weakens the distinctiveness of recommendation results. This phenomenon highlights the delicate balance between enhancing generalization and avoiding performance degradation.

Overall, moderate regularization plays a critical role in sequence modeling and attention optimization for advertising recommendations. Too small a coefficient fails to constrain learning, leading to insufficient generalization. Too large a coefficient restricts the effective use of feature information, reducing model performance. Therefore, in practice, it is important to select an appropriate regularization strength. This ensures that the system maintains stability while effectively capturing user interests and delivering precise advertising in complex environments.

This paper also gives the impact of the model hot update interval on the experimental results, and the experimental results are shown in Figure 3.

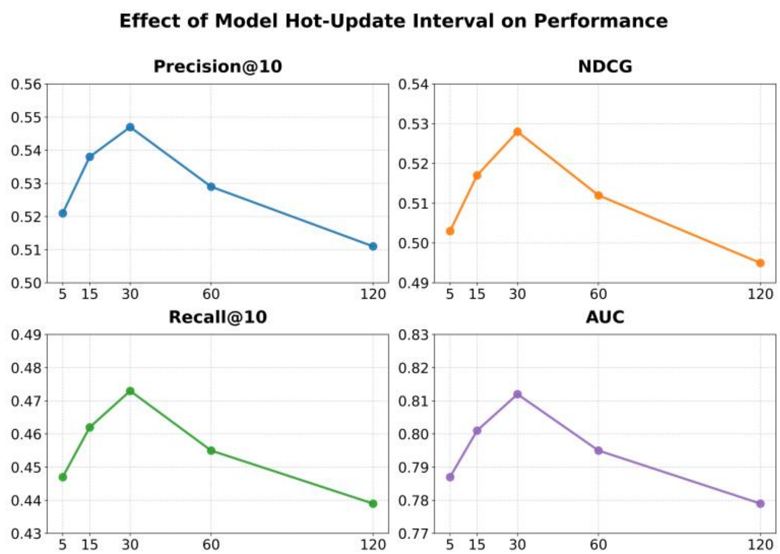

Figure 3. The impact of the model hot update interval on experimental results

The results show that the interval of model hot updates has a significant impact on overall recommendation performance. For Precision@10 and NDCG, when the update interval is short, between 15 and 30 minutes, the performance reaches its peak. When the interval extends to more than 60 minutes, performance decreases significantly. This indicates that frequent model updates help the system capture dynamic changes in user interests promptly, leading to better accuracy and ranking quality in recommendations.

The results for Recall@10 reveal a similar pattern. When the update interval is too long, the recall ability weakens, and the recommended results fail to cover enough relevant ads. In contrast, when the interval is very short, recall improves, but overly frequent updates may introduce noise. This reduces generalization on some samples. Therefore, a moderate update frequency can maintain coverage in the recommendation list while avoiding the performance fluctuations caused by overfitting.

For AUC, the trend is even more pronounced. As the update interval increases from 5 minutes to 30 minutes, the overall discriminative ability of the model improves steadily and reaches the best value at 30 minutes.

When the interval grows further, AUC declines, showing that the ability to distinguish between positive and negative samples weakens. This result further confirms that real-time recommendation systems need a balanced update frequency to ensure both discriminative power and system stability.

Considering all four metrics, the choice of hot update interval directly determines system performance in real-time environments. Very short intervals may increase system cost and accumulate noise, while very long intervals prevent the model from reflecting changes in user interests. The experimental results show that a reasonable update cycle achieves optimal performance in precision, recall, and discriminative ability. This provides important practical guidance for real-time advertising recommendations. It also shows that designing proper update strategies is essential to ensure both recommendation quality and user experience in dynamic environments.

This paper also provides a sensitivity analysis of the proportion of cold start users on Recall@10, and the experimental results are shown in Figure 4.

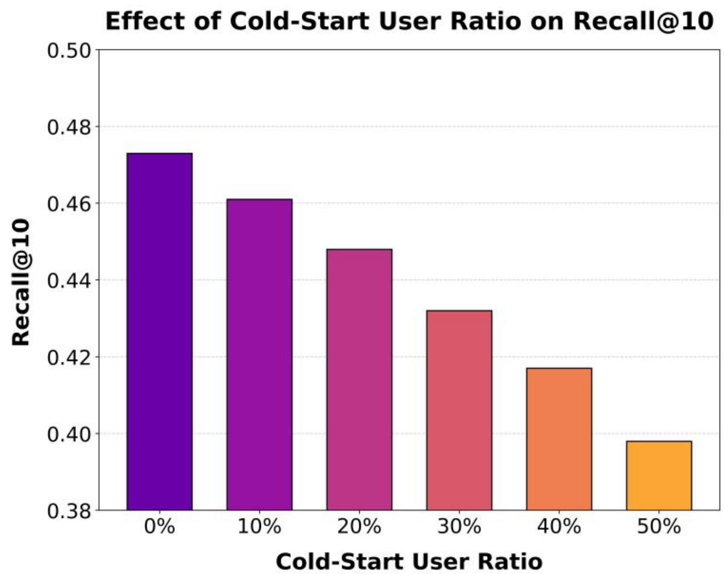

Figure 4. Sensitivity analysis of the proportion of cold start users to Recall@10

The experimental results show that an increase in the proportion of cold-start users has a clear negative impact on the Recall@10 metric. When the proportion of cold-start users is 0%, the model can fully use historical behavior sequences for modeling, which maintains a high recall rate. However, as the proportion of cold-start users increases, the available behavioral data decreases significantly. This limits the ability of the recommendation system to cover the full range of user interests, leading to a continuous decline in recall.

In the range of 10% to 30%, the decline of Recall@10 is relatively moderate. This indicates that the model can partly alleviate the cold-start problem through sequence modeling and attention optimization. The stability of performance in this stage shows that the system can still capture partial user preferences with limited contextual information, thus maintaining reasonable recommendation effectiveness. Yet as the proportion continues to rise, these compensatory mechanisms gradually fail, and the performance decline becomes more severe.

When the proportion of cold-start users exceeds 30%, the decrease in Recall@10 accelerates. At 50%, the recall rate drops to a low level. This means that when many users lack sufficient historical behavior sequences, the model cannot rely on existing mechanisms to infer interests. Both the coverage and diversity of recommendation results are seriously weakened. This highlights the critical constraint imposed by the cold-start problem on system performance. It also shows that relying only on behavioral sequence modeling has clear limitations in scenarios with a high proportion of cold-start users.

Overall, the results emphasize the need to design more effective strategies to mitigate cold-start effects in advertising recommendation systems. Possible approaches include combining static user profile features, contextual information, cross-user transfer learning, or graph-based modeling to enhance preference inference for cold-start users. Only when the cold-start problem is effectively addressed can sequence modeling and attention optimization achieve their full potential. This ensures that the system maintains stable recall performance and user experience in real applications.

This paper also presents a sensitivity analysis of the noise label injection rate to Precision@10, and the experimental results are shown in Figure 5.

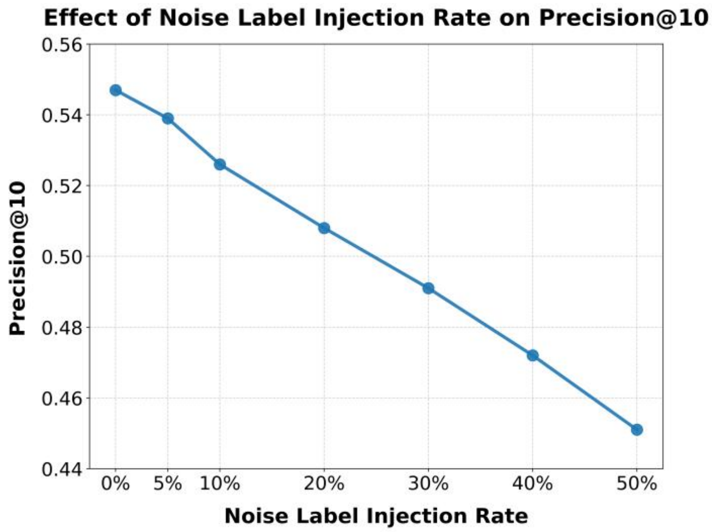

Figure 5. Sensitivity analysis of noise label injection rate to Precision@10

The experimental results show that an increase in the noise label injection rate has a continuous negative impact on Precision@10. When no noise labels are present, the model can fully rely on true supervision signals to achieve the highest recommendation accuracy. However, as the noise ratio increases, the supervision signals become contaminated. This leads to a gradual decline in the accuracy of user interest identification and ad matching. The trend clearly reflects the critical role of data quality in the performance of advertising recommendation systems.

At a low noise level, such as 5% to 10%, the decline of Precision@10 exists but remains moderate. This indicates that sequence modeling and attention mechanisms have some resistance to interference. The model can still maintain high recommendation accuracy even when part of the labels are inaccurate. This robustness benefits from the ability of the model to capture sequence dependencies and global context, which allows valuable behavioral patterns to be extracted despite the presence of noise.

When the noise level rises to 20% to 30%, Precision@10 decreases significantly. This shows that the robustness of the model is increasingly challenged. At this stage, biased labels reduce the proportion of relevant ads in the recommendation results. This directly weakens the usefulness of the short-list recommendation. Since advertising recommendation emphasizes precise matching, the interference of noisy labels is particularly severe. Both user experience and system effectiveness are affected. This indicates that recommendation systems in practice must rely strongly on data cleaning and quality control.

At a high noise level, such as 40% to 50%, Precision@10 falls to a low level. The model almost loses the ability to distinguish between highly relevant ads and less relevant ads. This highlights the destructive effect of label noise on sequence modeling and attention optimization. It also shows that relying only on supervised signals is clearly limited when data quality is insufficient. To ensure the effectiveness of advertising recommendations in real environments, stronger noise-robust mechanisms are necessary. Possible approaches include semi-supervised learning, self-supervised learning, or label correction strategies, which can fundamentally reduce the accuracy loss caused by data noise.

5. Conclusion

This paper investigated sequence modeling and attention optimization for real-time advertising recommendation, aiming to overcome the limitations of traditional static correlation methods. By introducing embedding representations and self-attention mechanisms, the proposed framework effectively captured both short-term fluctuations and long-term preferences in user behavior sequences. The integration of residual connections and regularization further enhanced stability, ensuring robust learning in highly dynamic and noisy recommendation environments.

Extensive experiments on the Taobao User Behavior dataset demonstrated the advantages of the method compared with several strong baselines. The results showed consistent improvements across Precision@10, Recall@10, NDCG, and AUC, highlighting the ability of the model to deliver both accurate and well-ranked recommendations. Sensitivity analyses on regularization strength, model update intervals, cold-start ratios, and label noise further confirmed the adaptability of the framework under diverse conditions.

The findings emphasize that sequential modeling and multi-level attention are indispensable for handling the complexity of real-world advertising scenarios. By explicitly modeling temporal dependencies and focusing on key fragments, the framework was able to mitigate challenges such as sparse signals, fast-changing interests, and noisy supervision. These contributions demonstrate the necessity of moving beyond static feature-based approaches toward dynamic and context-aware modeling.

In conclusion, the study provides both theoretical and practical contributions to the development of recommendation systems. The proposed framework enriches the methodology of sequence modeling with attention mechanisms and shows clear effectiveness in large-scale, real-time applications. Its improvements in recommendation accuracy, ranking quality, and robustness confirm its potential to serve as a reliable foundation for the next generation of intelligent advertising recommendation systems.

References

[1] Yang Y, Huang C, Xia L, et al. Multi-behavior hypergraph-enhanced transformer for sequential recommendation[C]//Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022: 2263-2274.

[2] Rashed A, Elsayed S, Schmidt-Thieme L. Context and attribute-aware sequential recommendation via cross-attention[C]//Proceedings of the 16th ACM conference on recommender systems. 2022: 71-80.

[3] Xie Y, Zhou P, Kim S. Decoupled side information fusion for sequential recommendation[C]//Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. 2022: 1611-1621.

[4] L. Xia, Y. Xu, C. Huang, P. Dai and L. Bo, "Graph meta network for multi-behavior recommendation," Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 757-766, 2021.

[5] J. Li, M. Wang, J. Li, J. Fu, X. Shen, J. Shang and J. McAuley, "Text is all you need: Learning language representations for sequential recommendation," Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1258-1267, 2023.

[6] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 639-648, 2020.

[7] W. C. Kang and J. McAuley, "Self-attentive sequential recommendation," Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), pp. 197-206, 2018.

[8] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian and X. Xie, "Self-supervised graph learning for recommendation," Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 726-735, 2021.

[9] X. Wang, X. He, Y. Cao, M. Liu and T. S. Chua, "KGAT: Knowledge graph attention network for recommendation," Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 950-958, 2019.

[10] W. C. Kang and J. McAuley, "Self-attentive sequential recommendation," Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), pp. 197-206, 2018.

[11] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian and X. Xie, "Self-supervised graph learning for recommendation," Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 726-735, 2021.