LatIA. 2025; 3:347
doi: 10.62486/latia2025347
ISSN: 3046-403X

ORIGINAL

Research on Intelligent Recommendation Algorithm of Short Videos Based on Graph Neural Network

Investigación sobre el algoritmo de recomendación inteligente de videos cortos basado en redes neuronales de grafos

Qiuye Guo[1,2], Sanghyun Kim[2]

[1]Guangzhou Vocational College of Technology & Business. Guangzhou City, Guangdong Province, 511442, P.R. China.
[2]Youngsan University, Yangsan Campus. 288 Junam-ro, Yangsan, Gyeongnam, 50510, Korea.

Cite as: Guo Q, Kim S. Research on Intelligent Recommendation Algorithm of Short Videos Based on Graph Neural Network. LatIA. 2025; 3:347. https://doi.org/10.62486/latia2025347

Submitted: 10-06-2024 **Revised:** 24-12-2024 **Accepted:** 19-07-2025 **Published:** 20-07-2025

Editor: PhD. Rubén González Vallejo

Corresponding Author: Sanghyun Kim

ABSTRACT

The rapid development of short video platforms has put forward higher requirements for the accuracy and personalization of content recommendation systems. In this paper, a short video recommendation algorithm based on Graph Neural Network (GNN) is studied, which improves the recommendation performance by fusing multimodal features such as video, audio, and text. The key technologies such as graph convolution neural network, graph attention network and graph pooling operator are analyzed, and a multimodal recommendation framework is constructed by combining self-supervised contrastive learning and local feature encoder to effectively deal with complex user-content interactions. In this paper, several algorithms are compared on TikTok and MovieLens datasets. The experimental results show that the SHL algorithm significantly improves the recommendation accuracy and user personalized satisfaction on TikTok and MovieLens datasets, which is generalizable.

Keywords: Graph Neural Network; Multimodal Recommendation; Self-Supervised Learning; Short Video; Personalization.

RESUMEN

El rápido desarrollo de las plataformas de vídeo corto ha planteado mayores exigencias para la precisión y personalización de los sistemas de recomendación de contenidos. En este trabajo se estudia un algoritmo de recomendación de video corto basado en redes neurongráficas (GNN) que mejora el rendimiento de la recomendación al fusionar características multimodcomo video, audio y texto. Se analizan las tecnologías clave como la red neuronal de convolución gráfica, la red de atención gráfica y el operador de grafo compartido, y se construye un marco de recomendación multimodal mediante la combinación del aprendizaje contrastivo autosupervisy el codilocal de características para hacer frente eficazmente a las complejas interacciones de contenido de usuario. En este trabajo, se comparan varios algoritmos en conjuntos de datos TikTok y MovieLens. Los resultados experimentales muestran que el algoritmo SHL mejora significativamente la precisión de la recomendación y la satisfacción personalizada del usuario en los conjuntos de datos TikTok y MovieLens, lo cual es generalizable.

Palabras clave: Red Neuronal Gráfica; Recomendación Multimodal; Autoaprendizaje; Video Corto; Personalización.

© 2025; Los autores. Este es un artículo en acceso abierto, distribuido bajo los términos de una licencia Creative Commons (https://creativecommons.org/licenses/by/4.0) que permite el uso, distribución y reproducción en cualquier medio siempre que la obra original sea correctamente citada

INTRODUCTION

As a core tool for analyzing non - Euclidean data, Graph Neural Networks (GNNs) have demonstrated significant advantages in fields such as social network analysis and bioinformatics. However, current research mainly focuses on modeling static graph structures, ignoring the dynamic evolution characteristics of network topologies in the real world. The existing approach of discretizing dynamic graphs into a sequence of time slices leads to the fragmentation of temporal information and the loss of long - term dependencies, making it difficult to meet the real-time decision-making requirements of highly dynamic tasks such as financial fraud detection and epidemic spread prediction. The essence of this limitation lies in the incompatibility between the static graph convolution mechanism and the dynamic system: the propagation paradigm based on a fixed adjacency matrix cannot represent the laws of edge weights decaying or mutating over time, and the state transitions between discrete snapshots further weaken the model's perception ability of the gradual change process. Due to the lack of explicit modeling of temporal causality, existing models often produce lagged responses and even misjudge key nodes.

Therefore, constructing a theoretical framework that can uniformly describe the evolution of graph structures and temporal dependencies has become a key breakthrough for improving the accuracy of dynamic system analysis. This study starts from the continuous - time representation of dynamic graphs and explores the coupling mechanism between graph topologies and temporal features, aiming to provide a more robust analytical basis for scenarios such as real time risk early warning and dynamic resource scheduling.

METHOD

Graph Convolution Neural Network

The spectral domain graph convolutional neural network mainly decomposes the features through the Laplacian matrix, while the spatial domain graph convolutional neural network mainly uses the aggregation analysis of the feature information of the adjacent graph nodes to define the convolution operation of the domain graph data features, and the space-based method redefines the graph convolution method by aggregating the feature information of the neighbor nodes.[1] In the training process of the model, each node of the above two methods will update the state of the previous node. Because the intermediate state of all nodes must be saved, the training efficiency is often not high. In order to solve this problem, some GCN methods have been extended in recent research work, such as GraphSAGE, which uses the combination of subgraph learning, MPNN, AGCN, MGCN, PinSage and Fast-GCN, which use information transfer. As well as graph convolution, there is also a lot of research on neural network optimization methods.[2] The structure of the convolutional neural network is shown in figure 1.

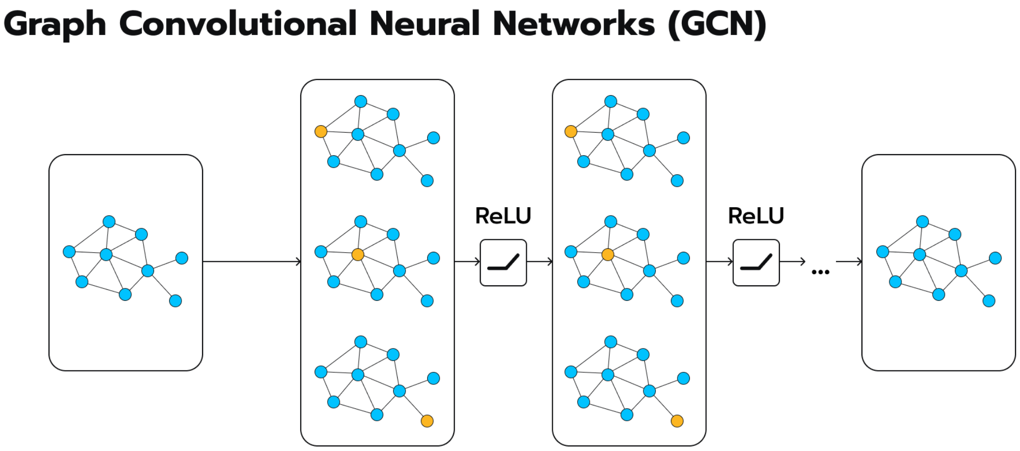


Figure 1. Structure of Graph Convolution Neural Network

Graph Attention Network

One of the greatest benefits of attention mechanism is that it can effectively expand the impact on the most critical part of the data, and now it has begun to be more and more widely used in a variety of research tasks based on time series analysis. This feature has also been gradually proved to be very advantageous for many tasks, such as machine language translation and natural language understanding.[3] Nowadays, the number of models with attention mechanism is also in a continuous growth trend, and the graph neural network model also benefits from this, which can effectively use attention in the process of information aggregation, and at the same time fuse the information output of each mode to form a random walk for multiple important information targets. The multimodal attention network is shown in figure 2.

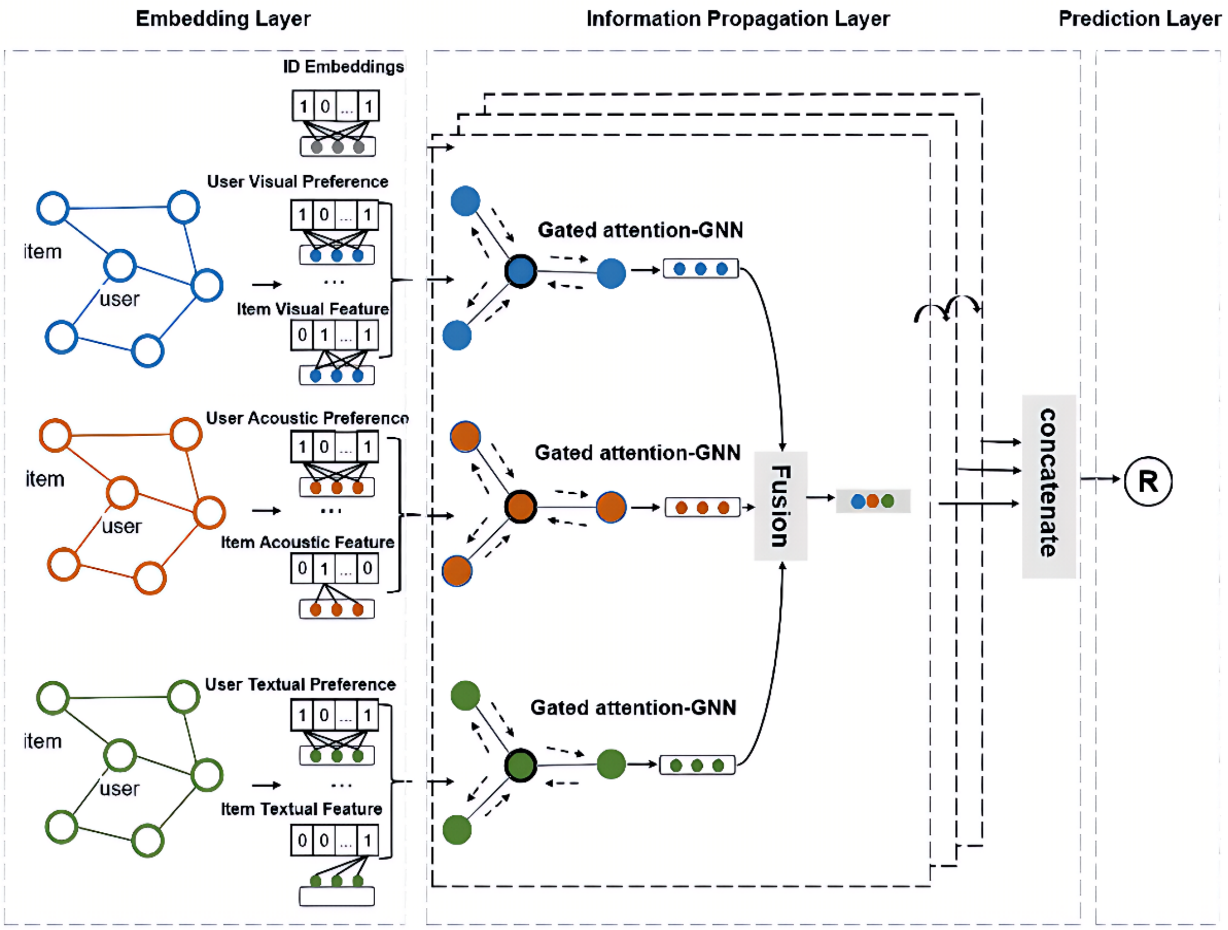


Figure 2. Multimodal attention network

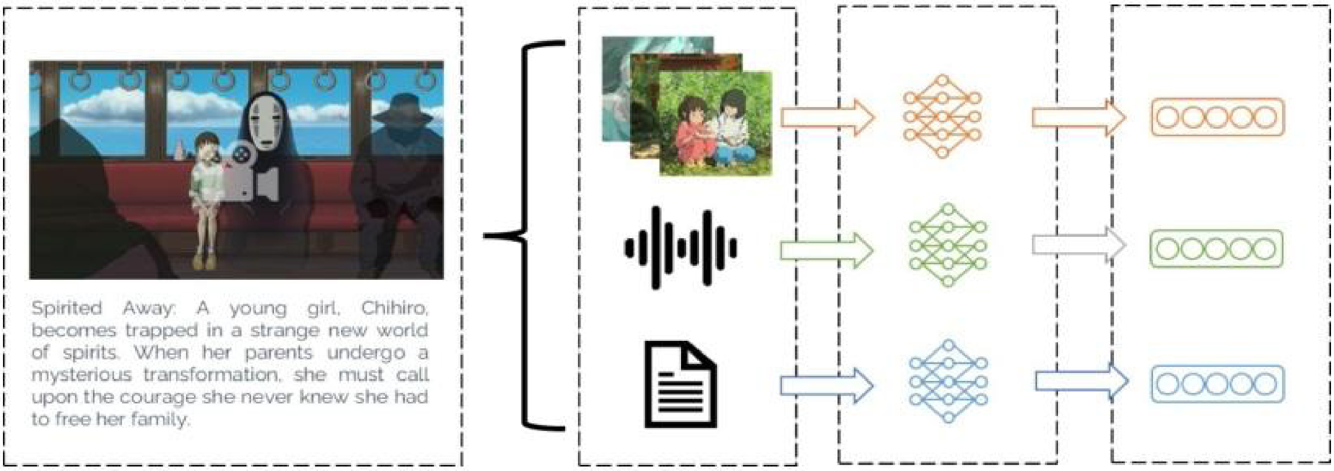
Graph Pooling Operator

At present, there are few graph pooling methods, but they can be roughly divided into several parts, namely, topological pooling, global pooling, and hierarchical pooling. The first topological pooling method is based on the ChebNet model of Graclus clustering algorithm.[4] The method it uses is not a neural network, but a graph coarsening algorithm that considers the structural characteristics of the graph. This spectral clustering algorithm can first be used to obtain a simplified graph structure, but the time complexity of its eigen-decomposition is very large, and it is necessary to find an alternative scheme. Global pooling is different from previous methods. Global pooling focuses on the overall features of the graph, and connects the convolutional layers of the graph to collect all the feature representations of the nodes in each layer. Global pooling can handle differently structured graphs because it brings together all the feature representations. Designed a graph classification framework using the Set2Set method, using GNN as a message passing scheme, and finally obtained the representation of the entire graph. A study proposed the SortPool layer, which sorts the nodes with different characteristics in the graph structure, and then delivers the sorted results to the next layer. A study proposed DiffPool, which is a differentiable graph pooling method that learns the assignment matrix in an end-to-end manner and eventually transforms into a clustering problem.[5] DiffPool algorithm divides the nodes of the original graph into multiple clusters, and then defines the clusters as the new nodes of the new graph, and then inputs the features of the new nodes into the next layer of the graph neural network.

Design of Short Video Intelligent Recommendation Algorithm Based on Graph Neural Network

Multi-Modal Feature Extraction

Multimodal feature extraction is an important step before model training. In this step, the model converts the multi-modal data into the feature vectors required for model training. In order to achieve this goal, three different pre-training models are used to transform video, audio and text data into feature vectors. The overall process is shown in figure 3.



Spirited Away: A young girl, Chihiro, becomes trapped in a strange new world of spirits. When her parents undergo a mysterious transformation, she must call upon the courage she never knew she had to free her family.

Figure 3. Multi-modal feature extraction process

When dealing with video data, this study uses the pre-trained ResNet50 model as the feature extraction model. This model can effectively extract rich feature information from video, including visual information such as color, texture and shape, and semantic information such as action and scene.[6] This information is critical to both the accuracy and robustness of the recommender system. Specifically, the model first abstracts the video, extracts the key frames of the video as the input parameters of the ResNet50 model, and then obtains a 2048-dimensional vector as the feature representation of the video after processing through the convolution layer and the fully connected layer of the model. This vector can effectively capture the key information in the video, and has good generalization performance. For audio data, this study uses Vggish model as a tool for feature extraction. The model is a VGG-like model pre-trained on YouTube AudioSet data, which can transform audio input into a high-level 128-dimensional feature vector with semantic meaning for the training task of downstream models. Because of its small size and fast computing speed, it is widely used in audio feature preprocessing and other tasks. When dealing with text data features, in order to save server resources and speed up feature extraction, this study uses a relatively simple Doc2Vec method.[7] By training the text representation model learned from the corpus, the method can transform the text into a vector with a fixed length, which is used in the subsequent recommendation model training. The method mainly learns a vector space by training the texts in the corpus, so that the texts in the same corpus have similar vector representations in the vector space. Doc2Vec can not only extract the semantic information of the text effectively, but also capture the context information of the text effectively.

Local Feature Encoder

In the local feature encoding module, the graph neural network is used as the basic framework, and the attention and gating mechanism are introduced to filter the local features respectively to ensure the reliability of the message transmission process. In order to obtain the learned node feature representation H, the whole local feature encoder is divided into three parts, namely, the feature representation of the node H, the ID feature representation of the node H, and the feature representation of the neighborhood nodes of the node H. Next, this chapter describes how to use the gated attention module to control the feature propagation of the neighborhood nodes of node H.

For different modes M, m is set as the feature representation of a certain mode, and the following formula is used to calculate the feature representation Nh of the neighboring nodes of the node H, and the specific calculation formula is as shown in (1):

$$e_{m,N_h} = R\left(\sum_{t\in N_h} \quad f_a(h,t) f_g(h,t) W_m^N e_{m,t}\right) \qquad (1)$$

Among, W_m^N represents a trainable parameter matrix, $e_{m,t}$ is a nodetIn the modem. The feature in represents,RCompared with the ReLU activation function, the LeakyReLU activation function can avoid the problem that neurons fall into a silent state when processing negative inputs, thus improving the efficiency and performance of network training to a certain extent. Functionf_a (h,t) and f_g (h,t) represent the attention and gating weight scores, respectively, through which data such as message content and the amount of information contained in the message can be effectively filtered and controlled during message delivery. Next, this chapter describes the functions individually.[8]

In the module of gating weight score calculation, the module adopts a unique gating component to precisely control the propagation of information flow. The core function of this gating component is to determine whether the feature information of neighboring nodes can be transferred to the current node. In this way, we can effectively control the direction of information flow and ensure the stability and performance of the network model.(9) In this method, the inner product of the central node and the surrounding nodes is used as the similarity of the two nodes, and it is used as the gating score. At the same time, in order to avoid too large values, the results need to be scaled appropriately. The specific formula is shown in (2):

$$f_g(h,t) = \delta\left(\frac{e_{m,h}^{\top} e_{m,t}}{\sqrt{d_t}}\right) \qquad (2)$$

Where H represents the center node, t represents the neighbor nodes of node H, d is the out-degree of node t, and δ is a Sigmoid function that maps the final score to the interval of 0 to 1 as the final weight score. In the module of attention weight score calculation, the module controls the passing weight of information flow by calculating the attention weight scores of the central node and the surrounding nodes. This design idea aims to achieve more accurate information filtering and information flow control to ensure that users can obtain the required content more efficiently and accurately when processing a large amount of information. The formula for calculating the weight is shown in (3,4):

$$f_a(h,t) = \left(W_{m,h} e_{m,h}\right)^{\top} tanh\left(W_{m,t} e_{m,t}\right) \qquad (3)$$

$$f_a(h,t) = \frac{exp^{f_a(h,t)}}{\sum_{t' \in N_h} \quad exp^{f_a(h,t')}} \qquad (4)$$

Here, m denotes the different mode spaces. The matrices Wmh and Wm are learnable transformation matrices. The tanh function is used as a nonlinear activation function, and the output value of the function is between -1 and 1, which can be conveniently normalized so that the weights between different features can be better compared and adjusted. In addition, the tanh function can be mapped to the entire range of real numbers compared to the Sigmoid function, which can only be mapped to the (0,1) interval, so its output range is wider.(10) And the mean of the Tanh function is close to 0, which helps to speed up the training of the model. After that, in order to determine the attention weight representing the relationship between two nodes, the model uses the inner product to calculate the weight score of two nodes, and then uses the sofmax function to normalize the attention weight of all neighbors. This allows the final attention score to distinguish the importance scores of different neighbors.

Finally, making use of the neighborInformation is propagated to update the representation of node H. Where the D characteristic of node H represents the, is considered as an anchor point across modes, as a highway to perform the propagation across modes. The specific formula is shown in (5,6):

$$\widetilde{e_{m,h}} = R\left(W_m^h e_{m,h}\right) + e_h \qquad (5)$$

$$e_{m,local} = R\left(W_m^1 e_{m,N_h}\right) + \widetilde{e_{m,h}} \qquad (6)$$

Self-Supervised Comparative Learning

Self-supervised contrastive learning can effectively enhance the robustness of the model. In this module, the module designs the contrastive learning component by computing the consistency between the local-based direct interaction relations and the hypergraph-based global implicit relations. This component leverages the self-discernment between the user and the entity to provide auxiliary supervisory signals from both local and global feature spaces. The fusion of multi-modal features has a very important impact on the results. Firstly, this study unifies the local and global features in each modality to generate a composite feature. e_m The method is shown in (7):

$$e_m = R\left(e_{m,local} + e_{m,global}\right) \qquad (7)$$

In the gate feature fusion strategy, the CONCAT function is first used to connect the features of different modes. Next, the weight information (representing the TORCH. MUL function) is calculated and obtained, and finally the final fusion feature data is obtained by using the weight information.

Intelligent Recommendation Module

In this experiment, Bayesian personalized ranking loss is used as the loss function Loss, whose goal is to maximize the score of positive samples and minimize the score of negative samples. The design of this loss function is helpful for the model to learn the difference between positive samples and negative samples better in the training process, so as to improve the accuracy and personalization of ranking. The specific calculation formula is as shown in (8,9):

$$y_{ui} = e_u^{\top} e_i \qquad (8)$$

$$Loss = \sum\nolimits_{(u,i,j)\in O} \quad - ln\left(\delta\left(y_{ui} - y_{uj}\right)\right) + L_{ssl} + \lambda \mid \theta \mid_2^2 \qquad (9)$$

R^+ represents a positive sample in the data set,R^-Representing the negative samples in the data set, the in-regularized weights,θIs the regularization parameter. In this study, three different pre-training models are selected to achieve accurate extraction of different modal features. Turning to the feature coding module on the right, you can see three different colors, which represent different modal features. Each modality has a shared trainable weight, which is designed to improve the training efficiency and flexibility of the model. In addition, the node N contains user feature nodes and entity feature nodes, and the lines between the nodes symbolize the interaction between different nodes. In particular, the model makes use of user ID as a cross-modal feature, making it a bridge for information transmission between different modalities. In the hypergraph learning module, the model innovatively introduces matrix factorization technology, which greatly reduces the number of parameters for model training, thus significantly improving the training speed. In addition, in order to enhance the robustness of the model, the model incorporates a self-supervised contrastive learning module, and improves the performance of the model through overall optimization.

Experimental Settings

An NVIDIA RTX 3090 graphics card was used in this experiment. Python is used as the programming language, and the above algorithms are implemented on the basis of Pytorch framework. In data set processing, each data set is randomly divided into three parts: training set (70 %), validation set (20 %) and test set (10 %). In terms of evaluation metrics, this experiment adopts three widely used metrics in the recommender system literature, namely Precision @ K, Recall @ K and Normalized Discounted Cumulative Gain NDCG @ K, where K is set to 10. In the training process, this experiment selects the learning rate from { 1e-3, 1e-4, 1e-5 }, uses the warm-up learning strategy, and uses Adam as the optimizer to optimize the model parameters.

Setting of Data Set

Two publicly available datasets, TikTok and MovieLens, are used in this experiment to evaluate the performance of the model. Table 1 gives some statistical information about the input data set.

Table 1. Introduction to Dataset

Datasets	#Interactions	#Items	#Users	Video	Audio	Text
TikTok	724329	27375	37024	128	128	128
MovieLens	1442586	5162	51852	2048	128	100

TikTok: this data set was first made available in the TikTok Data Mining Competition. The TikTok dataset includes short videos ranging in length from 3 to 15 seconds, as well as public text descriptions by video creators. The data was cleaned up before use, and the data items with missing and abnormal values were deleted. The data set contains the processed 128-dimensional video feature vector, 128-dimensional audio feature vector and desensitized text information of the TikTok project. The text information includes the title of the video and the brief introduction of the video.

Movie Lens: this dataset contains rating data for multiple movies provided by multiple users, as well as movie-specific metadata. In this work, we use the MovieLens-10M data set. In the completion of multi-modal feature data of movies, this paper uses crawler technology to grab the trailer data of related videos from YouTube. After that, ResNet50 network is used to extract visual features from video key frames. The trained VGGish network is used to learn audio features. Finally, the doc2vec algorithm is used to learn the text features, where the text contains the title and video introduction content.

RESULTS

In this experiment, Precision and Recall are used as evaluation indexes, and the following common baseline models are selected for comparison. The experimental results are shown in table 2.

Table 2. SHL comparison experiment results

Model	TikTok		MovieLens	
	Precision	Recall	Precision	Recall
ACF	0,091	0,201	0,099	0,39
GraphSAGE	0,107	0,225	0,105	0,41
NGCF	0,113	0,228	0,107	0,419
MMGCN	0,125	0,252	0,113	0,467
SASRec	0,15	0,31	0,115	0,495
SHL (OUR)	0,165	0,336	0,127	0,533

ACF: the ACF model utilizes an attention mechanism to capture implicit interactions between items and users to generate recommendations of multimedia content. GraphSAGE: GraphSAGE computes node embeddings from node feature information on invisible nodes. It learns the embedding by sampling and aggregating the feature information of the local neighbors of a node using a feature function. NGCF: The NGCF model implements a mechanism inspired by convolution in graph neural networks to explicitly model higher-order connectivity patterns. MMGCN: MMGCN constructs a separate bipartite graph for each modality to connect users and items in each modality, and then aggregates the representation information of these modalities to obtain the final user and item representation features for prediction. SASRec is a sequence recommendation algorithm based on self-attention mechanism, which can effectively capture the long-term and short-term dependencies in user behavior sequences, and is suitable for dealing with the sequence of user behavior. It is suitable for recommendation tasks that need to consider the order of user behavior, such as short video, e-commerce recommendation and so on.

DISCUSSION

The results of the proposed SHL learning method on TikTok and MovieLens10K datasets are better than those of the above baseline models. The present experiment attributes these improvements to the combination of features obtained by the local and global embedding layers, respectively. In this way, the model can capture the implicit relationships between nodes at a deeper level.

The model based on graph neural network performs better than traditional model based on collaborative filtering. These advances are attributed to the graph convolution layer, which can effectively capture the relationship between nodes and effectively combine the characteristics of node neighbors to improve the representation learning stage.

The hypergraph structure learning method added in this experiment can make up for the defects of the graph-based neural network and the collaborative filtering model, and can better find the deep-seated association information between nodes.

Compared with the existing literature, this paper has significant innovations and improvements in methodology. References [1,2] mainly analyze the phenomena and impacts of TikTok's recommendation algorithm, but do not delve into the recommendation algorithm based on graph neural networks. In this paper, however, a multi-modal recommendation framework is constructed, combining self-supervised contrastive learning and local feature encoders to effectively handle complex user-content interaction relationships. In addition, although [3] discusses multimedia recommendation systems, it does not involve the application of graph neural networks. By introducing graph neural network technology, this paper significantly improves the recommendation performance.

The experimental methods and procedures selected in this paper are fully reasonable. Key technologies such as graph convolutional neural networks, graph attention networks, and graph pooling operators are adopted, which can effectively handle complex user-content interactions. Through multi-modal feature extraction and the design of local feature encoders, video, audio, and text information can be efficiently integrated, providing a richer feature representation for the recommendation system.

Through reasonable method design and experimental procedures, this paper validates the effectiveness of the proposed method, conducts an in-depth comparison with the existing literature, and demonstrates its significant advantages in terms of recommendation accuracy and user-specific satisfaction.

CONCLUSIONS

Through the analysis of key technologies such as graph neural network, graph attention network and graph pooling operator, combined with multi-modal feature extraction and local feature encoder design, this paper constructs a recommendation framework that can efficiently integrate video, audio and text information. The experimental results show that the proposed algorithm has better performance than traditional algorithms in terms of recommendation accuracy and user personalized satisfaction. In addition, the self-supervised contrast learning and feature pyramid module are introduced to further improve the robustness and generalization ability of the model. Future research can further optimize the global feature encoder, explore a more refined hypergraph neural network model, and combine more diverse user behavior data to achieve a more accurate recommendation effect and provide a more personalized user experience for short video platforms.

BIBLIOGRAPHIC REFERENCES

1. Zhao Z. Analysis on the "Douyin (Tiktok) Mania" phenomenon based on recommendation algorithms[C]//E3S Web of Conferences. EDP Sciences, 2021, 235: 03029.

2. Zhang M, Liu Y. A commentary of TikTok recommendation algorithms in MIT Technology Review 2021[J]. Fundamental Research, 2021, 1(6): 846-847.

3. Deldjoo Y, Schedl M, Hidasi B, et al. Multimedia recommender systems: Algorithms and challenges[M]//Recommender systems handbook. New York, NY: Springer US, 2021: 973-1014.

4. Yanti D, Subagja A D, Nurhayati S, et al. Short Videos & Social Media Algorithms: Effective Communication in Tourism Marketing[J]. International Journal of Artificial Intelligence Research, 2024, 6(1.2).

5. Kirdemir B, Kready J, Mead E, et al. Examining video recommendation bias on YouTube[C]//International Workshop on Algorithmic Bias in Search and Recommendation. Cham: Springer International Publishing, 2021: 106-116.

6. Zhao H, Wagner C. How TikTok leads users to flow experience: investigating the effects of technology affordances with user experience level and video length as moderators[J]. Internet Research, 2022, 33(2): 820-849.

7. Khoo O. Picturing diversity: Netflix's inclusion strategy and the Netflix recommender algorithm (NRA)[J]. Television & New Media, 2023, 24(3): 281-297.

8. Fiallos A, Fiallos C, Figueroa S. Tiktok and education: Discovering knowledge through learning videos[C]//2021 Eighth International Conference on EDemocracy & EGovernment (ICEDEG). IEEE, 2021: 172-176.

9. Qin Y, Omar B, Musetti A. The addiction behavior of short-form video app TikTok: The information quality and system quality perspective[J]. Frontiers in Psychology, 2022, 13: 932805.

10. Zhan R, Pei C, Su Q, et al. Deconfounding duration bias in watch-time prediction for video recommendation[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 4472-4481.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Formal analysis: Qiuye Guo.
Research: Qiuye Guo.
Project management: Sanghyun Kim.
Supervision: Sanghyun Kim.
Drafting - original draft: Qiuye Guo.
Writing - proofreading and editing: Sanghyun Kim.