Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-Video Recommendation

Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu, *Fellow, IEEE*

***Abstract*—Micro-video recommendation has attracted extensive research attention with the increasing popularity of micro-video sharing platforms. Traditional approaches consider micro-video recommendation as a matching task and ignore the rich relationships among users and micro-videos from various modalities (e.g., visual, acoustic, and textual). Recently, GNN-based approaches show promising performance for the micro-video recommendation task. However, they mainly focus on the homogeneous graph which includes only one type of nodes or relations, and cannot be applied to the heterogeneous graph which consists of users, micro-videos, and related multi-modal information. In this paper, a novel Heterogeneous Hierarchical Feature Aggregation Network (HHFAN) is proposed for personalized micro-video recommendation. Our goal is to explore the highly complicated relationship information among users, micro-videos and related multi-modal information from a modality-aware Heterogeneous Information Graph (M-HIG), and thus generate high-quality user and micro-video embeddings for recommendation. The proposed model consists of two key components: (1) In data structure level, we build a heterogeneous graph and utilize a random walk based sampling strategy to sample neighbors for users and micro-videos. (2) In representation learning level, we design a hierarchical feature aggregation network including the intra- and inter-type feature aggregation networks to better capture the complex structure and rich semantic information in the heterogeneous graph. We evaluate our method on two real-world datasets and the results demonstrate that the proposed model outperforms the baseline methods.**

***Index Terms*—Heterogeneous graph, micro-video recommendation, multi-modal.**

I. Introduction

Now personalized recommendation is an important yet challenging task, which has attracted substantial attention in the past decade. Recently, people can easily shoot micro-videos or short-videos and share their daily stories on platforms like Tiktok[1] and Kwai,[2] which leads to the ever-increasing number of micro-videos. However, it is impossible for users to watch all available micro-videos due to time limitations. As a result, there is an enormous demand for micro-video recommendation to provide relevant information tailored to users' interests or preferences.

Most traditional approaches consider micro-video recommendation as a matching task [1], and can be solved by estimating the matching score based upon semantic representations of users and micro-videos [2]. Basically, there are three major research lines: (1) Feature combination based methods, such as Factorization Machine (FM) [3] and deep Factorization Machine (deepFM) [4], can learn feature combinations of users and micro-videos based on their simple feature information (e.g., locations, age, gender), for predicting interactions between users and micro-videos. (2) ID-based recommendation approaches, such as collaborative filtering (CF) [5] and low-rank factorization [6], only consider the interactions between users and micro-videos to obtain their corresponding representations, and rank micro-videos to the given user according to the feature similarity. (3) Multimedia micro-video recommendation methods mainly treat multi-modal information as a whole representation, reflecting similarity between different users and micro-videos by utilizing direct concatenation operation or attention mechanism, and incorporate them into a collaborative filtering framework, such as VBPR [7]. Although these algorithms show promising performance in the micro-video recommendation field, most of these methods only focus on simple features or multi-modal attribute information. In fact, there exists rich relationship information among users, micro-videos and related

Manuscript received October 17, 2020; revised January 26, 2021; accepted January 27, 2021. Date of publication February 22, 2021; date of current version February 8, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002804, the National Natural Science Foundation of China under Grants 62036012, 6207072426, 61720106006, 61572503, 61802405, 61872424, 61702509, 61832002, 61936005 and U1705262, in part by the Key Research Program of Frontier Sciences, CAS, under Grant QYZDJ-SSW-JSC039, and the K.C. Wong Education Foundation, in part by CCF-Tencent Open Fund, and in part by the Open Research Projects of Zhejiang Laboratory under Grant 2021KE0AB05. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wen-Huang Cheng. *(Corresponding author: Changsheng Xu.)*

Desheng Cai is with the Hefei University of Technology, Hefei 230009, China (e-mail: caidsml@gmail.com).

Shengsheng Qian and Quan Fang are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn; qfang@nlpr.ia.ac.cn).

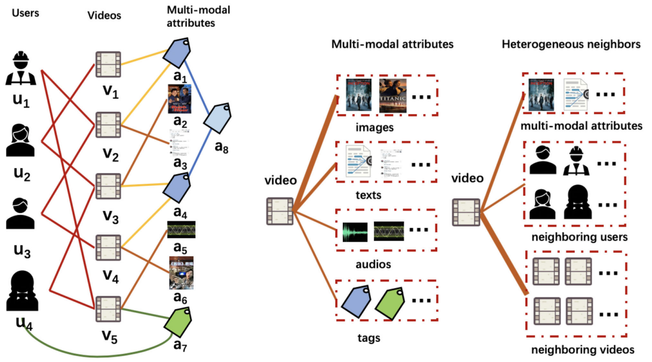
Changsheng Xu is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, with the , with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: csxu@nlpr.ia.ac.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3059508.

Digital Object Identifier 10.1109/TMM.2021.3059508

[1] [Online]. Available: https://www.tiktok.com/

[2] [Online]. Available: https://www.kwai.com/



(a) An example of heterogeneous graph

(b) Heterogeneous neighbors

Fig. 1. A heterogeneous graph and heterogeneous neighbors of users and micro-videos.

multi-modal information, which can be better exploited to obtain the high-quality representations. For example, users who have similar applications installed on their phones may be the same type of users. There is a close relationship between users with sports applications on their phones and users who watch sports-tagged micro-videos. Therefore, the relationship information should be considered to enrich the representations of users and micro-videos. Recently, graph representation learning approaches are emerging tools to pursue a meaningful vector representation for each node in graphs, which can effectively model users, micro-videos, their related multi-modal information and corresponding relationships. Graph Neural Networks (GNNs), such as GCN [8], GraphSAGE [9] and GAT [10], have shown impressive performance in aggregating feature information of neighboring nodes, which can be employed to micro-video recommendation. For example, MMGCN [11] is proposed to employ GCN on the modality-aware bipartite user-item graph to obtain better user representations based on item content information for micro-video recommendation. It seems reasonable to improve the performance of micro-video recommendation by exploiting additional relationship information among users and micro-videos. However, these existing methods still cannot effectively address the following challenges below.

1) In data structure level, most of current GNN-based micro-video recommendation methods focus on the homogeneous graph which includes only one type of nodes or links, and ignores rich relations existing among various kinds of heterogeneous data. In a real graph composed of users, micro-videos, and related multi-modal information, both nodes and links are heterogeneous. For example, as shown in Figure 1(a), this heterogeneous graph consists of three main types of nodes, user nodes, micro-video nodes and multi-modal attribute nodes. Here, multi-modal attribute nodes include text, image, audio, and tag nodes. In addition, there are many types of links representing the relationships among these different types of nodes, which are *Interaction relationships* (e.g., $u_1 - v_2, u_2 - v_1, u_3 - v_4$), *Co-occurrence relationships* (e.g., $v_2 - a_3$, $v_3 - a_2$, $v_5 - a_5$, $v_5 - a_7$, $u_4 - a_7$) and *Inclusion relationships* (e.g., $a_8 - a_1, a_8 - a_3$), respectively. We can find that the rich structural information in the heterogeneous graph can be utilized to reveal hidden inter-dependencies. Therefore, how to handle such complex structural information and preserve the diverse feature information simultaneously is a challenge in the heterogeneous graph.
2) In representation learning level, each node has multiple types of neighbors in the heterogeneous graph, which reflects the rich heterogeneous neighborhood information. For neighbors of same node types (i.e., intra-type nodes), their contents are homogeneous and very correlated. For example, as shown in Figure 1(b), image neighbors of each video usually are key frames extracted from the same corresponding video, which are highly correlated in content. For neighbors of different node types (i.e., inter-type nodes), they are heterogeneous and may contribute differently to the node embeddings. For example, in Figure 1(b), each micro-video node is not only associated with multi-modal attribute nodes, but also has connections to other user and micro-video nodes derived from various heterogeneous links. On the one hand, among the multi-modal attribute nodes, image and audio nodes should have more impact on the embedding generation of each micro-video node since visual and acoustic features are more specific and representative than textual features for micro-videos. On the other hand, for heterogeneous node types, user and micro-video neighbors of each micro-video node might be more important than multi-modal attribute neighbors for representation learning. Therefore, how to consider intra-type homogeneous node feature information for same type neighbors and different impacts of inter-type heterogeneous node feature information jointly to obtain a comprehensive representation for each user and micro-video is vitally important.

Based on the above discussions, we would like to consider these challenges in micro-video recommendations. In this paper, a new solution, named Heterogeneous Hierarchical Feature Aggregation Network (HHFAN), is proposed for personalized micro-video recommendation. Our approach consists of two key components: (1) In data structure level, we first build a heterogeneous graph, in which the node types include users, micro-videos and their corresponding multi-modal attributes, and the edge types include a variety of different co-occurrence or interaction relationships among different nodes. Then, we design a random walk based sampling strategy to sample neighbors for users and micro-videos of fixed size with strong correlations and group these sampled neighbors based on their node types for each user and micro-video. (2) In representation learning level, a novel hierarchical feature aggregation network is proposed to aggregate node feature information of these sampled heterogeneous neighboring groups of each user and micro-video. For nodes of the same type, we design an intra-type feature aggregation network to aggregate the intra-type homogeneous node feature information by employing the LSTM architecture for each generated group. For nodes of different types, we employ an inter-type feature aggregation network which consists of two different levels of attention, attribute-aware self-attention and neighbor-aware attention, to generate the final embeddings of users and micro-videos. Attribute-aware self-attention aims to learn the attention values among different multi-modal attribute nodes, and neighbor-aware attention focuses on learning the attention values of different type neighboring node groups (i.e., generated neighboring user embedding, aggregated attribute embedding and generated neighboring micro-video embedding). Finally, the proposed model predicts users' preferences by matching users and micro-video candidates with the respectively learned embeddings. In summary, the contributions of this work are as follows:

- We present a novel Heterogeneous Hierarchical Feature Aggregation Network (HHFAN) for personalized micro-video recommendation. Our model utilizes a heterogeneous graph to consider not only the information of users, micro-videos and their corresponding multi-modal content, but also the relationships among all types of nodes, which can enrich the representations of users and micro-videos and have the potential to generate better recommendation accuracy.
- Our proposed model utilizes a hierarchical feature aggregation network including both of the intra- and inter-type feature aggregation networks, which enables the learned node embeddings to better capture the complex structure and rich semantic information in the heterogeneous graph.
- We evaluate our method on the real-world datasets of Kwai, Tiktok and MovieLens, and the results demonstrate that the proposed model outperforms the baseline methods.

II. Related Work

This work is closely related to three fields, which are micro-video recommendation, graph neural networks and heterogeneous graph neural networks.

Micro-video recommendation: Recommendation is the most effective tool to alleviate the information overload problem on video-sharing platforms. Existing approaches to deal with video recommendation can be roughly grouped into three categories: collaborative filtering [12], [13], content-based filtering [14]–[17] and hybrid approaches [18]–[23]. However, most of them are proposed to deal with traditional videos (e.g., videos on YouTube), while little work has been conducted for micro-videos. In recent years, with the increasing popularity of micro-video sharing platforms (e.g. Kwai and Tiktok), micro-video recommendation [24]–[27] has attracted extensive research efforts to provide users with micro-videos in which they are interested. For example, Chen *et al.* [26] adopt a forward multi-head self-attention method with item-level and category-level attention module to model both behaviours and interests of users for micro-video recommendation. Recently, Liu *et al.* [28] use a stacked co-attention deep network to attend both user and video modalities and further sequentially model user and micro-video representations, which focus more on key features demonstrating users' hidden preferences. Although these approaches have achieved promising performance, they basically model users' preferences by utilizing users' historical interactions and the multi-modal content information of users and micro-videos. However, they largely ignore the rich relationship information among users, micro-videos, and their related multi-modal attribute information, such as micro-videos' tag trees and users' tag trees.

Graph neural networks: The key idea behind graph neural networks (GNNs) is to aggregate feature information from node's local neighbor information by neural networks. For instance, GCN [8], GraphSAGE [9] and Graph Attention Network [10] employ convolutional operator, LSTM aggregator and self-attention mechanism to aggregate neighbors' information, respectively. GNNs can be utilized to represent the embeddings of users and items in structural graphs for micro-video recommendations [11], [29]. For example, Wei *et al.* [11] propose a novel GCN-based framework, called MMGCN, leveraging information interchange between users and micro-videos to guide the representation learning in multiple modalities, and further model users' fine-grained preferences on micro-videos. However, the above GNNs are designed for embedding a homogeneous graph and thus fail to take advantage of the rich relations existing among various kinds of heterogeneous data.

Heterogeneous graph neural networks: Heterogeneous graphs (e.g. heterogeneous information networks) [30] can naturally model complex multiple types of objects and the rich relationships among them. There exit many heterogeneous graph based recommendation methods which rely on the path based similarity measures under different semantic meta-paths and further utilize a matrix factorization based on dual regularization framework for recommendation [31]–[34]. For example, Shi *et al.* [35] propose a novel heterogeneous network embedding based approach, called HERec, which can effectively integrate various kinds of embedding information based on different meta-paths in heterogeneous information network to enhance the personalized recommendation performance. Although the above methods can embed various heterogeneous graphs, they depend on the meta path selection, and may not fully consider the vital node feature information for recommendation. Recently, Zhang *et al.* [36] propose a heterogeneous graph neural network model (HetGNN) for heterogeneous node representation learning, which can jointly consider node heterogeneous content information, type-based neighbor aggregation, and heterogeneous type combination. However, HetGNN cannot fully utilize heterogeneous content information of nodes due to lacking of rich relationships among multi-modal content information of nodes. In order to explore the highly complicated relationships among users, micro-videos and their related multi-modal content information, our proposed model constructs a modality-aware heterogeneous information graph to enrich the representations of users and micro-videos for micro-video recommendation. In addition, a hierarchical feature aggregation network including both of the intra- and inter-type feature aggregation network, is proposed to better capture the complex structure and rich semantic information in the heterogeneous graph.

III. The Proposed Algorithm

A. Problem Statement

In this paper, we focus on the task of micro-video recommendation, in which all micro-videos are denoted as $V = \{v_1, v_2, \ldots, v_{|V|}\}$ and all users are denoted as $U =$

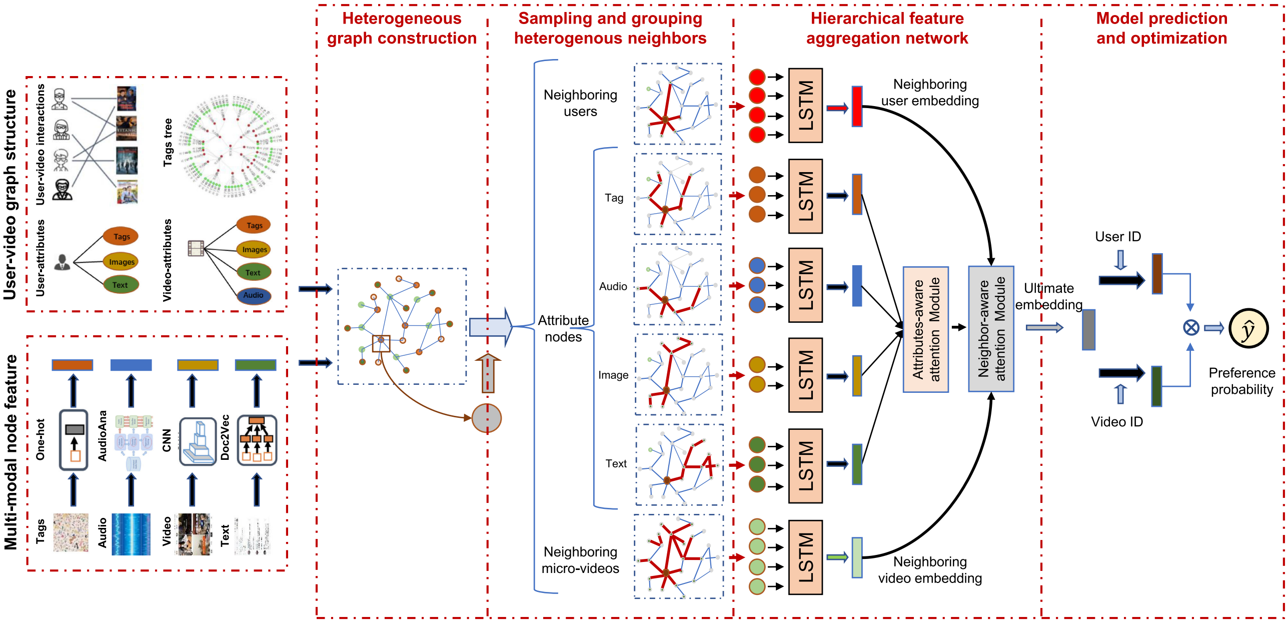


Fig. 2. HHFAN: Heterogeneous Hierarchical Feature Aggregation Network architecture.

$\{u_1, u_2, \ldots, u_{|U|}\}$. Each user is associated with some multi-modal attribute features denoted as U_{attrs} (e.g., gender, name, age, location, some tags of users, textual description, user picture): $U_{attrs} = \{u_1^a, u_2^a, \ldots, u_{|U_{attrs}|}^a\}$. Similarly, multi-modal attribute information of each micro-video is denoted as V_{attrs} (e.g., location, duration of micro-videos, some tags of videos, micro-video visual content, micro-video audio content, micro-video textual description, the title of micro-video): $V_{attrs} = \{v_1^a, v_2^a, \ldots, v_{|V_{attrs}|}^a\}$. Since both users and micro-videos have tag information, we use tag trees to describe the relationships between these tags. We define tag trees of users and micro-videos as a set $T = \{t^1, t^2, \ldots, v^{|T|}\}$ and a tag tree as $t^i = \{t_{jk}^i, j, k \in N+\}, 1 \leq i \leq |T|$, where t_{jk}^i represents the k-th tag node in the j-th layer of the i-th tag tree in the set T. We define a user-video interaction matrix as $I \in R^{|U| \times |V|}$, where the entry i_{uv} is defined from user's implicit feedback, $i_{uv} = 1$ indicates that the user u has clicked on the micro-video v and $i_{uv} = 0$ indicates that there is no click interaction between user u and micro-video v. Our purpose is to obtain corresponding high-quality representations of users and micro-videos given their multi-modal contents as well as relationships among them, and further calculate preference estimation scores to indicate users' preferences for personalized micro-video recommendation.

B. Overall Framework

To resolve the challenges described in Section 1, we present a novel Heterogeneous Hierarchical Feature Aggregation Network (HHFAN) for micro-video recommendation. Our framework, as shown in Figure 2, mainly consists of the following four components:

- **Heterogeneous Information Graph Construction**. We construct a heterogeneous graph to model users, micro-videos as well as their multi-modal attribute information based on the various relationships such as the co-occurrence information between users and their attributes, user-video interaction information and tag trees information.
- **Sampling and Grouping Heterogeneous Neighbors**. For each node in the constructed heterogeneous graph, we design and utilize a random walk based sampling strategy to sample fixed size correlated heterogeneous neighbors and group these sampled neighbors based upon their node types.
- **Hierarchical Feature Aggregation Network**. To aggregate node feature information of these sampled heterogeneous neighboring groups of each node, we design a novel hierarchical feature aggregation network with two steps: (1) Intra-type feature aggregation: For each generated group, we employ the LSTM architecture to aggregate the feature information of the neighboring nodes in the group, and obtain the aggregated type-based neighboring group embedding for each group. (2) Inter-type feature aggregation: After generating embeddings for each heterogeneous type-based neighboring group, attributes-aware self-attention network and neighbors-aware attention network are proposed to measure the different impacts of heterogeneous node types and obtain the ultimate node embedding for each node.
- **Model Prediction and Optimization**: The aggregated embeddings of users and candidate micro-videos are integrated through an inner product operation to infer the preference probability that indicates the degree of the user's preference for the candidate micro-videos. In addition, we use the Bayesian Personalized Ranking (BPR) [37] to optimize the HHFAN framework for predicting preference probabilities based on a designed generation strategy of training set.

C. Heterogeneous Information Graph Construction

In order to effectively and uniformly model information of users, micro-videos, their corresponding multi-modal attributes and various related relationships, we construct a modality-aware heterogeneous information graph (M-HIG), defined as $G = (N, E, TE_N, TE_E)$. In the graph M-HIG, N denotes multiple types of nodes and E represents relations between nodes, where $N = U \cup V \cup U_{attrs} \cup V_{attrs} \cup T$ and $E = I \cup E_{ua} \cup E_{va} \cup E_{ut} \cup E_{vt} \cup E_{tt}$. E_{ua}, E_{va}, E_{ut} and E_{vt} describe co-occurrence relationships of users and their associated multi-modal attributes, micro-videos and their related multi-modal attributes, users and tags as well as micro-videos and tags, respectively. E_{tt} represents inclusion relationships between tags in the tag trees T. TE_N represents the set of node types which consist of user nodes, micro-video nodes and attribute nodes: $TE_N = \{type^u, type^v, type^{attrs}\}$. Furthermore, $type^{attrs}$ is composed of multi-modal attribute node types, including image nodes, audio nodes, text nodes and tag nodes: $type^{attrs} = \{type^{image}, type^{audio}, type^{text}, type^{tag}\}$. TE_E is the set of relation types which contain the relation type of set E_{ua}, E_{va}, E_{ut}, E_{vt} and E_{tt}: $TE_E = \{type^i, type^{ua}, type^{va}, type^{ut}, type^{vt}, type^{tt}\}$.

D. Sampling and Grouping Heterogeneous Neighbors

Since the constructed M-HIG is heterogeneous, directly applying most existing GNNs to heterogeneous graphs may lead to two problems: (1) Heterogeneous neighbors may require different feature transformations to deal with different feature types and dimensions; (2) Existing GNNs approaches may not be able to effectively gather information between distant nodes or high-order relationship nodes.

To deal with these problems, we design a heterogeneous neighbor sampling strategy based on random walk, named Random Walk Sampling Strategy ($RWSS$). It contains two steps: (1) Starting a fixed-length random walk from node $v, \forall v \in V$. The walk iteratively travels to the neighbors of current node or returns to the starting node with a probability p. The random walker runs until it successfully collects a fixed number of nodes, denoted as $RWSS(v)$. Note that numbers of different types of nodes in $RWSS(v)$ are constrained to ensure that all node types are sampled for v. (2) Grouping different types of neighbors. For node v, we select top k_t nodes belonging to node type t from $RWSS(v)$ according to frequency and take them as the set of t-type correlated neighbors of node v.

Therefore, we can sample fixed size correlated heterogeneous neighbors and group these sampled neighbors based upon their node types for each node in the constructed heterogeneous graph. These generated groups can be divided into three main categories including multiple multi-modal attribute node groups, a neighboring user node group and a neighboring micro-video node group. Here, multi-modal attribute node groups can be further subdivided into an image node group, an audio node group, a text node group and a tag node group, which represent various associated multi-modal attributes of users and micro-videos respectively.

E. Hierarchical Feature Aggregation Network

Given a M-HIG G, we denote the feature representation of node $v \in V$ as $x_v \in R^{d_f \times 1}$, where d_f denotes node feature dimension. Note that x_v can be pre-trained or initialized using different techniques $w.r.t.$ different types of nodes. For example, we can utilize the doc2vec method [38] to pre-train nodes whose content type is text or employ CNN methods [39] to pre-train nodes whose content type is image. Here, we use a feature transformer layer denoted as $\mathcal{FC}_{\theta_x}$ to transform vectors of different types of nodes into a unified space. Note that $\mathcal{FC}_{\theta_x}$ can be a fully connected neural network with parameter θ_x, and is different $w.r.t.$ different types of nodes. Formally, the transferred node feature embedding of node v is computed as follows:

$$f(v) = \mathcal{FC}_{\theta_x}(x_v) \tag{1}$$

where $f(v) \in R^{d \times 1}$, and d is transferred embedding dimension.

To aggregate node feature embeddings transferred by $\mathcal{FC}_{\theta_x}$ of heterogeneous neighbors for each node, we next design a novel hierarchical feature aggregation network module which mainly includes two steps: (1) Intra-type feature aggregation; (2) Inter-type feature aggregation.

1) Intra-Type Feature Aggregation: After using the sampling strategy in section 3.4, we sample fixed size neighbor sets of different node types for each node. We denote the t-type sampled neighbor set of $v \in V$ as $N_t(v)$, and employ a neural network to aggregate node feature embeddings of $v_{ner} \in N_t(v)$. Formally, the aggregated t-type neighbors' embedding for v is formulated as follows:

$$f^t(v) = Aggregator^t_{v_{ner} \in N_t(v)} \{f(v_{ner})\} \tag{2}$$

where $f^t(v) \in R^{d \times 1}$, d denotes aggregated t-type neighbors' embedding dimension, $f(v_{ner})$ is the transferred node feature embedding of node v_{ner}, and $Aggregator^t$ is the t-type neighbors aggregator function. In this work, inspired by GraphSAGE [9], we choose the LSTM as the $Aggregator^t$ function. Thus we re-formulate $f^t(v)$ as follows:

$$f^t(v) = \frac{\sum_{v_{ner} \in N_t(v)} LSTM\{f(v_{ner})\}}{|N_t(v)|} \tag{3}$$

We use LSTM to aggregate transferred node feature embeddings of t-type neighbors and conduct the average operation over all hidden states to generate the aggregated t-type neighbors' embedding.

2) Inter-Type Feature Aggregation: The previous step generates aggregated t-type neighbors' embedding for node v. The M-HIG G has three main node types, user nodes $type^u$, micro-video nodes $type^v$ and multi-modal attribute node types $type^{attrs}$. Different types of neighbors in the heterogeneous graph may contribute differently to the node embeddings. In order to combine these embeddings into the final representation of node v by considering the impacts of these embeddings on node v, we employ an inter-type feature aggregation network which consists of two different levels of attention, attribute-aware self-attention and neighbor-aware attention, to generate the final embeddings of users and micro-videos.

Specifically, attribute-aware self-attention aims to learn the attention values between different multi-modal attribute nodes, and neighbor-aware attention aims to learn the attention values of different neighboring node groups for heterogeneous node types. Based on the learned attention values, our model can consider the complex structure information of the heterogeneous graph to learn better node representations.

Attributes-aware self-attention network The previous step generates $|type^{attrs}|$ (i.e., four) aggregated multi-modal attribute neighboring embeddings, denoted as $\mathcal{E}_v^{type^{attrs}} \in R^{|type^{attrs}| \times d}$ for node v. In order to combine these embeddigns into a final representation for all multi-modal attribute groups according to their importance, we employ a self-attention technique. Following [40], we define the attention operation with the residual connection as follows:

$$Attention(X) = softmax\left(\frac{XW^Q(XW^K)^T}{\sqrt{d_K}}\right) XW^V + X \tag{4}$$

where $W^Q \in R^{d\times d}, W^K \in R^{d\times d}, W^V \in R^{d\times d}$ are different transform matrices of the input X. Then we compute the self-attention based multi-modal attributes combined embedding as follows:

$$\mathcal{E}_v^{attrs} = \sum Attention(\mathcal{E}_v^{type^{attrs}}) \tag{5}$$

where $\mathcal{E}_v^{attrs} \in R^{d\times 1}$, and d denotes the final multi-modal attribute embedding dimension.

Neighbors-aware attention network After the previous two steps, $|TE_N|$ (i.e., three) aggregated neighbor embeddings are generated for node v, denoted as $\mathcal{E}_v^{TE_N} \in R^{|TE_N| \times d}$. In order to combine these embeddigns into the final representation of node v by considering the different impacts of these embeddings on node v, we employ an attention technique. Thus, we compute the attention based combined final embedding of node v as follows:

$$\mathcal{E}_v = \sum \alpha^{v,i} \mathcal{E}_{v,i}^{TE_N} \tag{6}$$

$$\alpha^{v,i} = \frac{exp(LeakyRelU(u^T[v \bigoplus \mathcal{E}_{v,i}^{T_V}]))}{\sum exp(LeakyRelU(u^T[v \bigoplus \mathcal{E}_v^{T_V}]))} \tag{7}$$

where $\mathcal{E}_v \in R^{d\times 1}$ is the combined final embedding of node v, $\alpha^{v,*}$ indicates the importance of different embeddings and $u^T \in R^{2\,d\times 1}$ is the attention parameter.

Note that in this paper, we use the same dimension d in the transferred node feature embedding, aggregated t-type neighbor embedding, and combined final embedding for node v to make model adjustment easier.

F. Model Prediction and Optimization

After feature aggregation operation, embeddings $\mathcal{E} \in R^{|N| \times d} (d \ll |V|)$ of users and micro-videos are generated. For each user, the main goal of HHFAN framework is to predict a list of micro-videos that a user may prefer. Therefore, given a user u and a candidate micro-video v, we compute a predicted preference probability $\hat{y}$ that indicates how much user u prefers micro-video v, based upon their learned representations. Formally, we define the preference probability calculation as follows:

$$\hat{y}_{uv} = \sigma({\mathcal{E}_u}^T \otimes \mathcal{E}_v) \tag{8}$$

where $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$ is the sigmoid function, $\mathcal{E}_u$ and $\mathcal{E}_v$ are learned representations of user u and candidate micro-video v respectively, and $\hat{y}_{uv}$ is the predicted preference probability for user u and candidate micro-video v. In order to effectively generate node representations and precisely calcualte preference probabilities, we apply the Bayesian Personalized Ranking (BPR) [37] method for optimizing HHFAN framework. In particular, we model a triplet of one user and two micro-videos, in which one of the micro-videos is preferred and the other one is not, formally as,

$$\mathcal{R} = \{< u, v, v' >\} \tag{9}$$

where $\mathcal{R}$ is a set of triples for training.

To consider the diversity of nodes in M-HIGs in the training stage, we design a generation strategy of training set $\mathcal{R}$. It contains two steps:

1) **Sampling neighbors:** For each user u in graph we use the random walk to sample a set of associated micro-videos $A'(u)$ which must satisfy the following two conditions: a) The number of nodes for each node type $t \in TE_N$ is fixed; b) Distance $dist(u, v_a) \leq \tau, \forall u_a \in A'(u), \tau \in \mathbb{N}^+$;
2) **Generation:** we generate the set of triples $< u, v, v' >$ as follows: a) For each user u, we generate $|A'(v)|$ pairs, denoted as a set $\{< u, v_k > | v_k \in A'(u), k = 1, \ldots, |A'(u)|\}$; b) For each pair $< u, v_k >$, we sample M nodes, $(v'_1, v'_2, \ldots, v'_M,)$, and then generate M triples, denoted as a set $\{< u, v_k, v'_m >, m = 1, \ldots, M\}$;

There are two advantages for this design as follows: (1) The main function of the sample neighbor step is to measure users' preference for videos by the distance function, $dist() \leq \tau, \tau \in \mathbb{N}^+$. The motivation behind this is that a user may not only prefer a set of videos that are directly linked to him, but may also prefer a set of videos that are indirectly linked to him. For example, if user A and user B have a lot of videos they prefer in common, then user A is likely to prefer other videos that user B prefers, even if user A is not directly connected to those videos ($dist() > 1$). This design actually benefits from the idea of collaborative filtering [5] in recommendation systems and makes the proposed model more generalized. (2) The purpose of the generation step is to reasonably search a number of negative videos for each user's preferred video, which is used to train the proposed model for generating more robust user and video representations and further achieving better recommendation results.

Note that, for each iteration in the training phase, the above two steps are performed to generate $|A'(u)| * M$ training triples for a user u. Furthermore, it is assumed that the user prefers the observed micro-video rather than the unobserved one. Therefore, we can reformulate the loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = \sum_{<u,v,v'>\in\mathcal{R}} - \ln \sigma({\mathcal{E}_u}^T \otimes \mathcal{E}_v - {\mathcal{E}_u}^T \otimes \mathcal{E}_{v'}) + \lambda ||\theta||_2^2 \tag{10}$$

TABLE I
STATISTICS OF DATASETS

Dataset	User	micro-video	Interaction	Sparsity
Kwai	169,878	310,681	775,834,643	98.53%
Tiktok	3,656	7,085	1,253,112	95.51%
MovieLens	6040	3706	1,000,209	95.53%

where $\sigma(x)$ is the sigmoid function, $\mathcal{E}_u$, $\mathcal{E}_v$ and $\mathcal{E}_{v'}$ are corresponding learned representations of user u, micro-video v and micro-video v', and λ and θ represent the regularization weight and the parameters of the model, respectively.

IV. EXPERIMENTAL RESULTS

A. Datasets

We conduct experiments on three real-world datasets: Kwai dataset and Tiktok dataset. We summarize the statistics of datasets in Table I, and more details are described as follows.

- **Kwai dataset:**[3] It is a real-world dataset obtained from the Kwai platform, which consists of users, micro-videos and all of their relevant data including interactive data for users, users' attributes, video' attributes, and additional relational data, such as category information of users and videos.
- **Tiktok dataset:**[4] It is released by Tiktok, a micro-video sharing platform that allows users to create and share micro-videos with duration of 3-15 seconds. This dataset consists of users, micro-videos and their interactions (e.g., click). Here, the micro-video features in each modality are extracted and published without providing the raw data.
- **MovieLens (MLs) dataset:**[5] This movie rating dataset has been widely used to evaluate collaborative filtering algorithms. We used the version containing one million ratings, where each user has at least 20 ratings. To this end, we transformed it into implicit data, where each entry is marked as 0 or 1 indicating whether the user has rated the item.

B. Baselines

We employ the following state-of-the-art micro-video recommendation algorithms as baselines.

- **FM$_{HIG}$:** It is a context-aware factorization machine [41] algorithm, which is able to utilize various kinds of auxiliary information. In our experiments, we extract heterogeneous information as context features and incorporate them into the factorization machine for micro-video recommendation.
- **DeepFM:** DeepFM [4] is a factorization-machine based neural network which contains an FM component and a deep component to combine the power of factorization machines and deep learning for feature representations.
- **NeuMF:** Neural Matrix Factorization [42] is a popular collaborative filtering based method, which can combine generalized matrix factorization model and deep learning to explore behavior information.
- **GraphSAGE:** GraphSAGE [9] is based on the general inductive framework that leverages node feature information to update node representations for the previously unseen data. In particular, it considers the structure information as well as the distribution of node features in the neighborhood.
- **PinSAGE:** PinSAGE [29] is a GCN-based algorithm, which combines efficient random walks and graph convolutions to generate embeddings of nodes.
- **MMGCN:** MMGCN [11] is a multimodal GCN recommendation framework, which generates modal-specific representations of users and micro-videos by graph convolution operation to capture user preferences.
- **SemRec:** SemRec [34] is a recommendation method utilized in weighted heterogeneous information graphs, which uses a weighted meta-path strategy to obtain users' different preferences on paths. It only mines the path semantic for recommendation without using multi-modal video features.
- **HeRec:** HeRec [35] is a heterogeneous information graphs based ranking method that merges different meta-path strategies to learn node embeddings for recommendations.
- **HAN:** HAN [43] embeds HINs by first converting an HIN to several homogeneous sub-networks through pre-defined meta-paths and then applying graph attention networks.
- **HetGNN:** HetGNN [36] is a heterogeneous graph neural network model for heterogeneous node representation learning, which can jointly consider node heterogeneous content information, type-based neighbor aggregation, and heterogeneous type combination.

For fair comparisons for the baselines, all baselines take external information, such as category information, into consideration for micro-video recommendation. Traditional models, FM, DeepFM and NMF, take all attribute information of users and videos, including external category information, as input. And for graph-based baselines, GraphSAGE, PinSAGE, MMGCN, SemRec, HeRec, HAN and HetGNN, they all take a modality-aware heterogeneous information graph (M-HIG), which treats attribute information (e.g. category information) as graph nodes, as input.

C. Evaluation Metrics and Parameter Settings

For each dataset, we randomly hold 20% of micro-videos associated with each user to form the testing set, and assign other micro-videos to the training set. We employ cross-validation with grid-search to tune all hyper-parameters in the training set. For testing set, we pair each observed user-item pair with 1000 unobserved micro-videos that the user has not interacted with before. We evaluate our method and other compared ones on micro-video recommendation by using four popular metrics [11], Precision at top K (P@K), Recall at top K (R@K), Normalized Discounted Cumulative Gain at top k (NDCG@k)

[3] [Online]. Available: https://www.kwai.com/

[4] [Online]. Available: http://ai-lab-challenge.bytedance.com/tce/vc/

[5] [Online]. Available: http://grouplens.org/datasets/movielens/1m/

TABLE II
PERFORMANCE COMPARISONS BETWEEN OUR MODEL AND THE BASELINES ON THREE DATASETS. (K=10, **TRAINING _ RATE=0.8**)

Datasets	Metrics	FM_{HIG}	DeepFM	NeuMF	GraphSage	PinSage	MMGCN	SemRec	HeRec	HAN	HetGNN	HHFAN
Kwai	Pre	0.1132	0.1263	0.1724	0.2304	0.3414	0.3819	0.3763	0.3801	0.3763	0.3821	**0.4233**
	Rec	0.2539	0.2697	0.3074	0.3767	0.4137	0.4102	0.3982	0.3897	0.3701	0.4101	**0.4566**
	NDCG	0.3305	0.3426	0.3904	0.4395	0.4222	0.4601	0.4637	0.4731	0.4543	0.4602	**0.4923**
	AUC	0.6140	0.6480	0.7270	0.7390	0.7320	0.7509	0.7691	0.7424	0.7511	0.7608	**0.7944**
Tiktok	Pre	0.1345	0.1163	0.1924	0.2111	0.2801	—-	0.3521	0.3629	0.3543	0.3722	**0.3823**
	Rec	0.1928	0.2445	0.2933	0.3709	0.3899	—-	0.3798	0.3514	0.3601	0.3801	**0.4491**
	NDCG	0.3215	0.3433	0.3705	0.4404	0.4198	—-	0.4402	0.4401	0.4204	0.4451	**0.4693**
	AUC	0.5925	0.6108	0.6527	0.7091	0.7023	—-	0.6901	0.6978	0.7302	0.7309	**0.7599**
MLs	Pre	0.1244	0.1364	0.2001	0.1999	0.2801	—-	0.3521	0.3529	0.3531	0.3491	**0.3701**
	Rec	0.1801	0.2301	0.2801	0.3501	0.3899	—-	0.3798	0.3514	0.3421	0.3821	**0.4173**
	NDCG	0.3441	0.3778	0.3765	0.4100	0.4121	—-	0.4237	0.4331	0.4241	0.4302	**0.4555**
	AUC	0.6029	0.6599	0.6783	0.6771	0.6701	—-	0.6888	0.6999	0.7107	0.7223	**0.7457**

TABLE III
PERFORMANCE COMPARISONS BETWEEN OUR MODEL AND THE BASELINES ON THREE DATASETS. (K=10, **TRAINING _ RATE=0.6**)

Datasets	Metrics	FM_{HIG}	DeepFM	NeuMF	GraphSage	PinSage	MMGCN	SemRec	HeRec	HAN	HetGNN	HHFAN
Kwai	Pre	0.1599	0.1771	0.1919	0.3019	0.3319	0.3777	0.3743	0.3712	0.3792	0.3892	**0.4038**
	Rec	0.2524	0.2801	0.3301	0.3569	0.4029	0.4012	0.3972	0.3828	0.3741	0.4010	**0.4211**
	NDCG	0.3225	0.3329	0.3991	0.4010	0.4333	0.4311	0.4235	0.4201	0.4301	0.4331	**0.4638**
	AUC	0.6595	0.6881	0.7091	0.6891	0.6969	0.7202	0.7129	0.7007	0.7172	0.7199	**0.74931**
Tiktok	Pre	0.1321	0.1103	0.1962	0.2119	0.2551	—-	0.3213	0.3199	0.3223	0.3201	**0.3446**
	Rec	0.1727	0.2312	0.2817	0.3109	0.3592	—-	0.3797	0.3779	0.3762	0.3771	**0.4033**
	NDCG	0.3023	0.3441	0.3515	0.3505	0.3552	—-	0.4402	0.4401	0.4204	0.4451	**0.4641**
	AUC	0.5925	0.6108	0.6527	0.7091	0.7023	—-	0.6901	0.6978	0.7309	0.7309	**0.7272**
MLs	Pre	0.1247	0.1227	0.1990	0.2108	0.2991	—-	0.3401	0.3222	0.3331	0.3351	**0.3448**
	Rec	0.1887	0.2219	0.2881	0.3011	0.3899	—-	0.3791	0.3691	0.3531	0.3881	**0.4012**
	NDCG	0.3441	0.3908	0.3265	0.4001	0.4021	—-	0.4097	0.4191	0.4141	0.4202	**0.4209**
	AUC	0.6343	0.6599	0.6621	0.6719	0.6867	—-	0.6818	0.6819	0.6807	0.6823	**0.7221**

and AUC. Here we set K = 10 and report the average scores in the testing set.

$$\mathbf{Precision@K} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\left(\frac{\sum_{j=1}^{K} pc(j, g(u))}{K}\right) \tag{11}$$

where j denotes the predict item, $g(u)$ represents the ground-truth items associated with user u, $pc(j, g(u))$ is an indicator function that returns 1 if j is in $g(u)$ and 0 otherwise, and K is the truncation level.

$$\mathbf{AUC} = \frac{1}{\mathcal{U}}\sum_{u\in\mathcal{U}}\frac{1}{|J||J'|}\sum_{j\in|J|}\sum_{j\in|J'|}\delta(p_{u,j} > p_{u,j'}) \tag{12}$$

where J denotes the positive samples set, J' means negative, $\delta(p_{u,j} > p_{u,j'})$ is an indicator function which returns 1 if $(p_{u,j} > p_{u,j'})$ is true, and 0 otherwise, and $p_{u,j}$ is the predicted probability that a user $u \in \mathcal{U}$ may act on i in the test set. The higher the AUC value, the better the ranking performance. The floor of AUC from random guess is 0.5 and the best result is 1.

To train our proposed model, we use a Gaussian distribution to randomly initialize the model parameters with a mean of 0 and standard deviation of 0.1, use the LeakyReLU as the activation function, and optimize our model through mini-batch Adaptive Moment Estimation (Adam) [44]. We search the batch size in {128, 256, 512}, the learning rate in {0.0001, 0.0005, 0.001.0.005, 0.01} and the regularizer in {0, 0.0001, 0.0001, 0.001, 0.01, 0.1}. As the findings are consistent across the dimensions of latent vectors, if not otherwise specified, we only show the result of 200, a relatively large number that returns good performance.

D. Performances Comparison

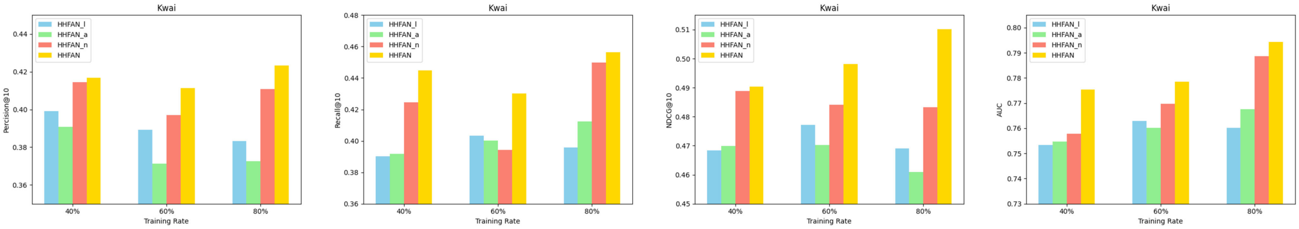
The comparative results are summarized in Table II, III and IV, with 80%, 60%, 40% of data being used as training sets. The proposed HHFAN model consistently beats all the baselines in terms of all metrics on three datasets as shown in Table II, III and IV, verifying the effectiveness of our proposed model for micro-video recommendation. From the results, we have the following observations:

- NeuMF has better performance than the feature-based models, such as FM and DeepFM, on the three datasets. The main reason may be that NeuMF model can make good use of multi-modal and interaction data.
- Both GraphSAGE and PinSAGE outperform the NeuMF model on the three datasets. This shows that graph convolution operations can not only capture the local structure information but also learn the distribution of neighbors' features for each node, which can boost the expressiveness of representations for micro-video recommendation.

TABLE IV
PERFORMANCE COMPARISONS BETWEEN OUR MODEL AND THE BASELINES ON THREE DATASETS. (K=10, **TRAINING _ RATE=0.4**)

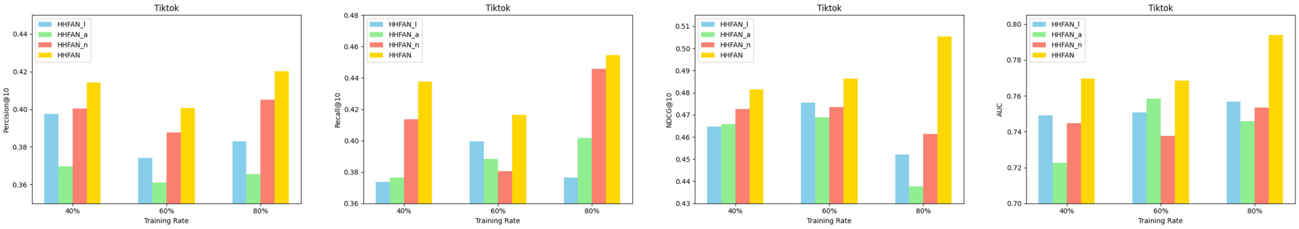
Datasets	Metrics	FM_{HIG}	DeepFM	NeuMF	GraphSage	PinSage	MMGCN	SemRec	HeRec	HAN	HetGNN	HHFAN
Kwai	Pre	0.1139	0.1264	0.1663	0.2103	0.3332	0.3719	0.3752	0.3822	0.3732	0.3891	**0.4138**
	Rec	0.2536	0.2597	0.3104	0.3553	0.4019	0.4102	0.3982	0.3897	0.3701	0.4101	**0.4411**
	NDCG	0.3215	0.3316	0.3331	0.4002	0.4391	0.4601	0.4637	0.4731	0.4543	0.4602	**0.4598**
	AUC	0.6110	0.6555	0.7301	0.7111	0.7119	0.7298	0.7119	0.3242	0.7371	0.7438	**0.7601**
Tiktok	Pre	0.1001	0.1101	0.1925	0.2118	0.2501	—-	0.3451	0.3429	0.3523	0.3501	**0.3746**
	Rec	0.1627	0.2312	0.2818	0.3706	0.3897	—-	0.3798	0.3614	0.3841	0.3991	**0.4233**
	NDCG	0.3023	0.3441	0.3515	0.3505	0.3552	—-	0.4202	0.4201	0.4214	0.4351	**0.4541**
	AUC	0.67915	0.6506	0.6927	0.6879	0.7023	—-	0.6907	0.6998	0.7171	0.7201	**0.7372**
MLs	Pre	0.1447	0.1727	0.1999	0.2808	0.2971	—-	0.3301	0.3122	0.3221	0.3331	**0.3668**
	Rec	0.1987	0.2819	0.2981	0.3211	0.3479	—-	0.3751	0.3691	0.3631	0.3841	**0.4131**
	NDCG	0.3551	0.3708	0.3665	0.3301	0.4063	—-	0.4099	0.4161	0.4131	0.4221	**0.4491**
	AUC	0.6293	0.6590	0.6321	0.6419	0.6567	—-	0.6618	0.6719	0.6707	0.6723	**0.7101**

- MMGCN achieves better performance than GraphSAGE and PinSAGE. It is because MMGCN employs GCN on the modality-aware bipartite user-item graph instead of unifying multi-modal features, which can obtain more fine-grained modal-specific representations of users and micro-videos. Compared to MMGCN, our model performs better for the following reasons: (1) Our model considers hidden, rich and heterogeneous relationships among multi-modal attributes to enrich the representations of users and micro-videos. However, MMGCN leverages Graph Convolution Network (GCN) framework on a user item bipartite graph in each modality to learn modal-specific representations of users and micro-videos and ignores these hidden, rich and heterogeneous relationships among multi-modal attributes. (2) Our model utilizes a novel hierarchical attention network, which can consider different impacts of heterogeneous node feature information on embedding generation, to obtain a final comprehensive representation for each user and micro-video However, MMGCN combines modal-specific representations for generating final representations of users and micro-videos without considering their modal-specific impacts. It just leverages two combination functions, which are the concatenation combination function and the element-wise combination function.
- Meta-path based baselines on heterogeneous information graphs, SemRec, HeRec and HAN, achieve better results than other baselines over three datasets in most cases. This indicates that the heterogeneous graph can model rich relationship information among users and micro-videos, and meta-paths can capture these information, which is very important for micro-video recommendation. Compared with meta-path based models, our method obtains meaningful improvements due to the hierarchical heterogeneous neighbor aggregation mechanism. The aggregation mechanism can learn the importance of different sampled neighbors to generate better user/video representations. The importance is automatically obtained with an end-to-end hierarchical attention network and therefore no prior knowledge is required. Instead, the meta-path baselines require the manual design of meta-path schemas for neighbor aggregation. The design of meta-path schemas highly relies on expert knowledge and improper meta-path schemas may introduce noisy information into the user/video representations.
- HetGNN, which is based on message propagation mechanism on heterogeneous information graphs, achieves better results than meta-path based recommendation baselines over three datasets in all cases. This indicates that HetGNN can more effectively leverage and aggregate heterogeneous information for node representation learning, having better performance on micro-video recommendations. Compared with HetGNN, the state-of-the-art heterogeneous graph representation learning method, our model performs better. The main reason is that our model considers hidden, rich and heterogeneous relationships among multi-modal attributes to enrich the representations of users and micro-videos for the micro-video recommendation. To achieve this, our model absorbs multi-modal attribute data into heterogeneous graph nodes, instead of just considering users and micro-videos as nodes, and further utilizes the heterogeneous and rich relationships among these multi-modal attributes as edges in the heterogeneous graph. More specifically, our model has the following advantages: (1) The sampling and grouping module of our model searches and samples relevant heterogeneous neighboring nodes of each node based on rich relationships among multi-modal attributes and further makes our model generate a more robust and comprehensive representation of each user and each micro-video. However, HetGNN just utilizes user-item interactions and ignores these relationships among multi-modal attribute data. (2) The feature aggregation module of our model uses a novel hierarchical attention network, which consists of attribute-aware self-attention and neighbor-aware attention, to take the importance of multi-modal attributes and different neighboring node types into consideration simultaneously for node representation learning. However, HetGNN does not consider the importance among different multi-modal attributes for representation learning of each node.



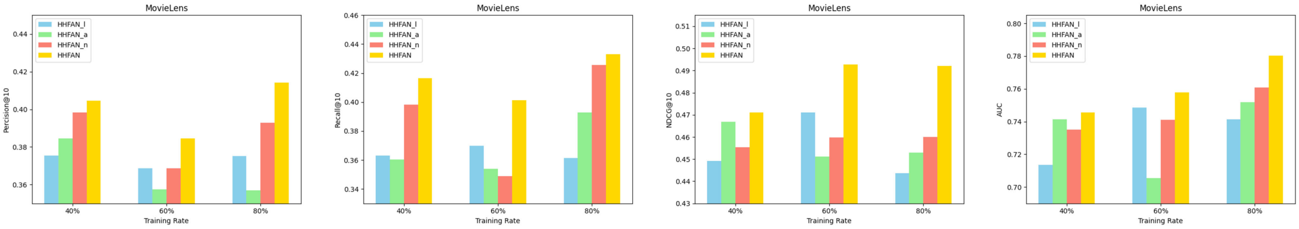
(a) Precision (b) Recall (c) NDCG (d) AUC

Fig. 3. Performance of Precision@10, Recall@10, NDCG@10 and AUC w.r.t. the rate of training data on the **Kwai dataset**.



(a) Precision (b) Recall (c) NDCG (d) AUC

Fig. 4. Performance of Precision@10, Recall@10, NDCG@10 and AUC w.r.t. the rate of training data on the **Tiktok dataset**.



(a) Precision (b) Recall (c) NDCG (d) AUC

Fig. 5. Performance of Precision@10, Recall@10, NDCG@10 and AUC w.r.t. the rate of training data on the **MLs dataset**.

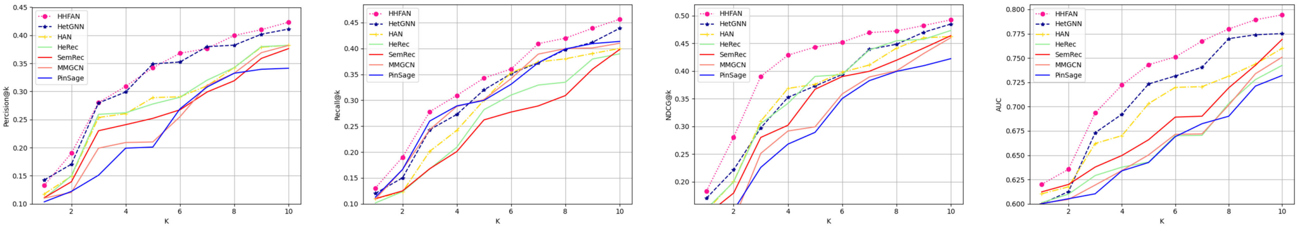
E. Ablation Study

We compare the performance of HHFAN and all variants of HHFAN, and report the results of Precision@10, Recall@10, NDCG@10 and AUC w.r.t. the rate of training data on three dataset in Fig. 3, Fig. 4 and Fig. 5.:

Impact of The LSTM Architecture: The LSTM architecture can not only model the information of the input data itself, but also consider the relational information existing in the input data. To verify the effectiveness of the LSTM network, we design a variant called **HHFAN**$\neg l$, which removes the LSTM in the intra-type feature aggregation module and only uses the average operation over transferred node feature embeddings in groups to generate the aggregated corresponding neighbors' embedding. HHFAN has better performance than HHFAN$\neg l$ in all cases, demonstrating that the LSTM architecture is better than the "shallow" encoding for aggregating type-based neighboring group embedding. These results may be due to the fact that the LSTM structure can not only integrate the information of the same type nodes, but also absorb the relationship information between the same type nodes for aggregating type-based neighboring group embedding.

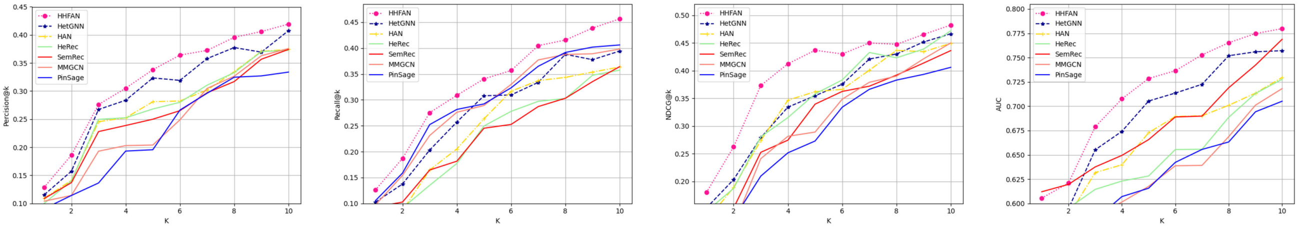
Impact of The Attribute-aware Self Attention Network: The attribute-aware self attention network aims to learn the attention values among different multi-modal attribute nodes for generating the aggregated attribute embedding for each node. To verify the effectiveness of the Attribute-aware self attention network, we design a variant called **HHFAN**$\neg a$, which removes the attributes-aware self-attention in the inter-type feature aggregation module and assigns the same importance to multi-modal attribute nodes. HHFAN outperforms HHFAN$\neg a$, showing that the importance of different modal attributes (e.g., visual, acoustic, and textual) can be better obtained by the attributes-aware self-attention module. These results are reasonable because the attribute information of different modes contributes differently to each node representation. For example, image and audio attribute information should have more influence on the embedding generation of each micro-video node and representations of user nodes are more dependent on textual attribute information.

Impact of The Neighbor-aware Attention Network: The neighbor-aware attention network focuses on learning the attention values of different type neighboring node groups (i.e., generated neighboring user embedding, aggregated attribute embedding and generated neighboring micro-video embedding) for generating the final embedding of each node in the M-HIG. To validate that the the neighbor-aware attention network can benefit the proposed model, we introduce it into HHFAN to build a variant called **HHFAN**$\neg n$, which removes the neighbors-aware



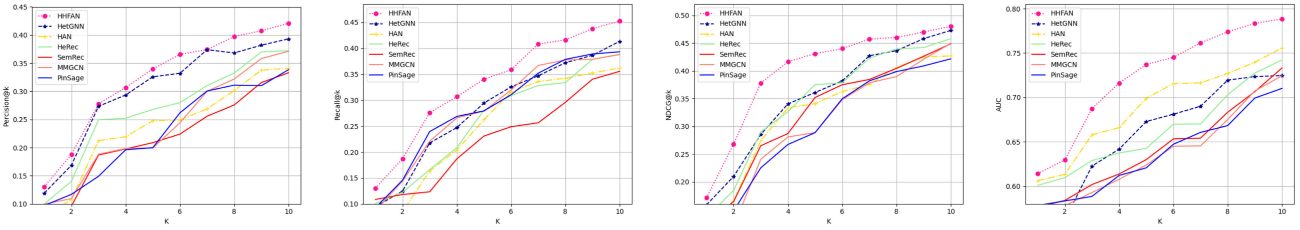
(a) Precision@K (b) Recall@K (c) NDCG@K (d) AUC

Fig. 6. Evaluation of Top-K item recommendation where K ranges from 1 to 10 on the **Kwai dataset**. training_rate=0.8.



(a) Precision@K (b) Recall@K (c) NDCG@K (d) AUC

Fig. 7. Evaluation of Top-K item recommendation where K ranges from 1 to 10 on the **Tiktok dataset**. training_rate=0.8.



(a) Precision@K (b) Recall@K (c) NDCG@K (d) AUC

Fig. 8. Evaluation of Top-K item recommendation where K ranges from 1 to 10 on the **MovieLens dataset**. training_rate=0.8.

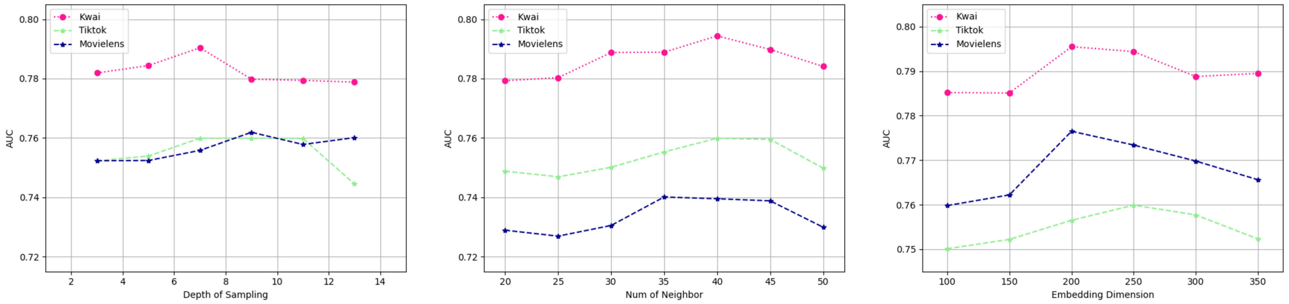
attention in the inter-type feature aggregation module and assigns the same importance to neighboring node groups. In the results, the performance of HHFAN is superior to HHFAN$\neg n$, indicating that the neighbors-aware attention module can measure the influence of different type neighboring node groups (i.e., generated neighboring user and micro-video embedding, aggregated attribute embedding) for generating final node embeddings. The reason for these results may be that user and micro-video neighbors are more relevant to the micro-video recommendation scenario and are more vital for representation learning.

F. Hyper-Parameters Sensitivity

We conduct experiments to analyze the impacts of four key parameters of the proposed HHFAN including the ranking position K, the depth of sampling, the aggregated embedding dimension and the number of sampled neighbors set for users and micro-videos.

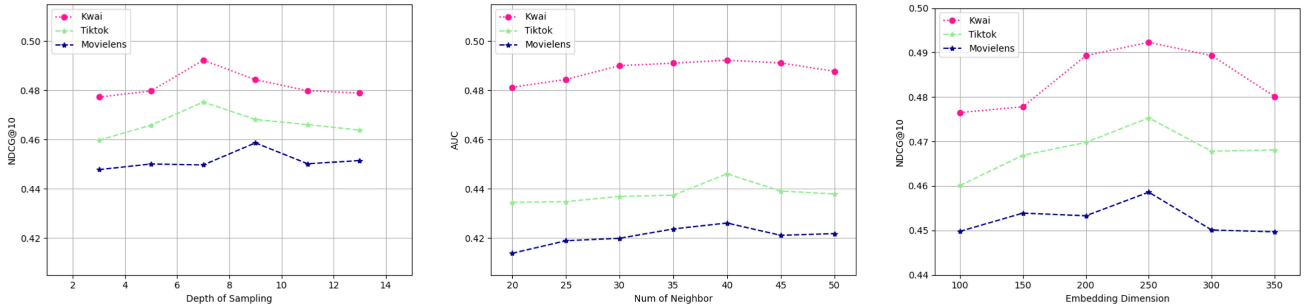
The Ranking Position K: Fig. 6, Fig. 7 and Fig. 8 show the performance of Top-K recommended lists on three datasets respectively, where the ranking position K ranges from 1 to 10. It can be seen that HHFAN has consistent improvements over other methods across positions, indicating the necessity of modeling heterogeneous information, including heterogeneous nodes and heterogeneous relationships, as well as the better capacity of representation leaning of our model. And compared with other methods, our method increases more stably with the increase of K, implying that our model is more stable in micro-video recommendation. For better display effect, the K value in the experiment is set to 10.

The Depth of Sampling, The Aggregated Embedding Dimension and The Number of Sampled Neighbors Set: Fig. 9 and Fig. 10 show the corresponding results of AUC and NDCG@10 on three datasets, respectively. From the results, we have the following observations: (1) When the depth of sampling for each node varies from 2 to 14, results of AUC and NDCG@10 rise slowly. However, as the sampling depth increases further, the performance becomes worse. The reason may be that uncorrelated ("noise") neighbors are involved. (2) When the number of neighbors for each node varies from 20 to 50, results of AUC and NDCG@10 rise steadily to the optimum. In addition, when the number of neighbors exceeds a certain value, the performance begins to decline slowly. The reason may also be the influence of noise data, resulting in poor performance of the model. (3) When the aggregated embedding dimension



(a) Depth of Sampling (b) Num of Neighbor (c) Embedding Dimension

Fig. 9. Performance of AUC of HHFAN in term of different parameters on three datasets. k=10, training_rate=0.8.



(a) Depth of Sampling (b) Num of Neighbor (c) Embedding Dimension

Fig. 10. Performance of NDCG@10 of HHFAN in term of different parameters on three datasets. k=10, training_rate=0.8.

d of each node varies from 100 to 350, AUC and NDCG@10 are basically increasing. However, with the further increase of d, the performance declines slowly, which may be due to overfitting. The experimental results of these three parameters are relatively stable. In order to better display the experimental results, we set the three parameters to the values that achieve the best experimental results. In our experiment, the specific values of the depth of sampling, the aggregated embedding dimension and the number of sampled neighbors are set to 6250 and 35, respectively.

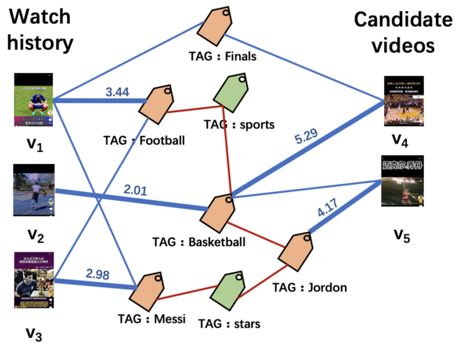


Fig. 11. Visualization of randomly sampled videos of a user w.r.t. a set of candidate videos.

G. Case Study

To intuitively demonstrate the effectiveness of the HHFAN in exploring the highly complicated relationships, such as relationships among related multi-modal attributes of users and micro-videos, we visualize some micro-videos, related tags and relationships of the M-HIG, constructed on the Kwai dataset. We randomly sample a user with three historical watched videos (v1,v2,v3) and the candidate videos (v4, v5) for recommendation. For each video, we calculate the (unnormalized) relevance probability between videos and tags and highlight paths between videos and their highest probability tags with corresponding probabilities, as shown in Figure 11. We observe that HHFAN can associate videos with their most relevant tags, such as “v1”-”TAG:Football,” and further make the learned representations of watched videos and candidate videos closer via several paths, such as “v2”-”TAG:Basketball”-”v4”. HHFAN can automatically mine these important paths, which can be used to explain the recommendation result. Therefore, recommendations tell users what videos they might want to click and why they click on them, which can help improve the performance of micro-video recommendation.

V. Conclusion

In this paper, we investigate the problem of personalized micro-video recommendation. We argue that most GNN-based approaches only learn models in the homogeneous graph and ignore rich relations existing among various kinds of heterogeneous data in the real micro-video recommendation graph. A novel Heterogeneous Hierarchical Feature Aggregation Network (HHFAN) is proposed to explore the highly complicated relationships among users, micro-videos and related multi-modal information, and generate high-quality user and micro-video embeddings for recommendation. By constructing a modality-aware heterogeneous information graph, our approach can model the highly complicated additional relationships among users, micro-videos and their related multi-modal content information. In addition, a hierarchical feature aggregation network including both of the intra- and inter-type feature aggregation network, is proposed to better capture the complex structure and rich semantic information to enrich the representations of users and micro-videos. Experimental results on two micro-video datasets show that our algorithm outperforms existing methods in the micro-video recommendation task. In the future, we will investigate the influence of users' social network information or knowledge graph information combined with the current heterogeneous graph on the model.

References

[1] J. Xu, X. He, and H. Li, "Deep learning for matching in search and recommendation," in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 1365–1368.

[2] X. He *et al.*, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[3] S. Rendle, "Factorization machines," in *Proc. 10th IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 995–1000.

[4] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: A factorization-machine based neural network for CTR prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1725–1731.

[5] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, pp. 421 425:1–421 425: 19, 2009.

[6] X. He, H. Zhang, M. Kan, and T. Chua, "Fast matrix factorization for online recommendation with implicit feedback," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Jul. 2016, pp. 549–558.

[7] R. He and J. J. McAuley, "VBPR: Visual bayesian personalized ranking from implicit feedback," in *Proc. 13th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 144–150.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, Apr. 2017, pp. 1–14.

[9] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 1024–1034.

[10] P. Velickovic *et al.*, "Graph attention networks," in *6th Int. Conf. Learn. Representations, Conf. Track Proc.*, Apr./May 2018.

[11] Y. Wei *et al.*, "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia, MM*, Oct. 2019, pp. 1437–1445.

[12] N. Koenigstein and U. Paquet, "Xbox movies recommendations: Variational bayes matrix factorization with embedded feature selection," in *Proc. 7th ACM Conf. Recommender Syst.*, RecSys '13, Oct. 2013, pp. 129–136.

[13] Y. Huang *et al.*, "Real-time video recommendation exploration," in *Proc. Int. Conf. Manage. Data, SIGMOD Conf.* Jun./Jul. 2016, pp. 35–46.

[14] T. Mei, B. Yang, X. Hua, and S. Li, "Contextual video recommendation by multimodal relevance and user feedback," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, pp. 10:1–10:24, 2011.

[15] S. E. Shepstone, Z. Tan, and S. H. Jensen, "Using audio-derived affective offset to enhance TV recommendation," *IEEE Trans. Multim.*, vol. 16, no. 7, pp. 1999–2010, Nov. 2014.

[16] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 191–198.

[17] T. Han *et al.*, "Dancelets mining for video recommendation based on dance styles," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 712–724, Apr. 2017.

[18] Z. Wang *et al.*, "Joint social and content recommendation for user-generated videos in online social network," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 698–709, Apr. 2013.

[19] Q. Huang, B. Chen, J. Wang, and T. Mei, "Personalized video recommendation through graph propagation," *TOMCCAP*, vol. 10, no. 4, pp. 32:1–32:17, 2014.

[20] A. Ferracani, D. Pezzatini, M. Bertini, and A. D. Bimbo, "Item-based video recommendation: An hybrid approach considering human factors," in *Proc. ACM Int. Conf. Multimedia Retrieval*, Jun. 2016, pp. 351–354.

[21] P. Zhou, Y. Zhou, D. Wu, and H. Jin, "Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1217–1229, Jun. 2016.

[22] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 609–618, Mar. 2017.

[23] Z. Zhao *et al.*, "Social-aware movie recommendation via multimodal network learning," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 430–440, Feb. 2018.

[24] J. Chen *et al.*, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 335–344.

[25] L. Huang and B. Luo, "Personalized micro-video recommendation via hierarchical user interest modeling," in *Proc. Adv. Multimedia Inf. Process. - PCM 18th Pacific-Rim Conf. Multimedia, Revised Sel. Papers, Part I, ser. Lecture Notes Comput. Sci.*, Sep. vol. 10735, 2017, pp. 564–574.

[26] X. Chen *et al.*, "Temporal hierarchical attention at category- and item-level for micro-video click-through prediction," in *Proc. ACM Multimedia Conf. Multimedia Conf., MM*, Oct. 2018, pp. 1146–1153.

[27] J. Ma *et al.*, "LGA: Latent genre aware micro-video recommendation on social media," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 2991–3008, 2018.

[28] S. Liu, Z. Chen, H. Liu, and X. Hu, "User-video co-attention network for personalized micro-video recommendation," in *Proc. World Wide Web Conf.*, May 2019, pp. 3020–3026.

[29] R. Ying *et al.*, "Graph convolutional neural networks for web-scale recommender systems," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2018, pp. 974–983.

[30] C. Shi, B. Hu, W. X. Zhao, and P. S. Yu, "Heterogeneous information network embedding for recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 357–370, Feb. 2019.

[31] C. Shi, J. Liu, F. Zhuang, P. S. Yu, and B. Wu, "Integrating heterogeneous information via flexible regularization framework for recommendation," *Knowl. Inf. Syst.*, vol. 49, no. 3, pp. 835–859, 2016.

[32] J. Zheng *et al.*, "Recommendation in heterogeneous information network via dual similarity regularization," *Int. J. Data Sci. Anal.*, vol. 3, no. 1, pp. 35–48, 2017.

[33] B. Hu, C. Shi, W. X. Zhao, and P. S. Yu, "Leveraging meta-path based context for top- n recommendation with a neural co-attention model," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2018, pp. 1531–1540.

[34] C. Shi *et al.*, "Semrec: A personalized semantic recommendation method based on weighted heterogeneous information networks," *World Wide Web*, vol. 22, no. 1, pp. 153–184, 2019.

[35] C. Shi, B. Hu, W. X. Zhao, and P. S. Yu, "Heterogeneous information network embedding for recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 357–370, Feb. 2019.

[36] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2019, pp. 793–803.

[37] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, Jun. 2009, pp. 452–461.

[38] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31th Int. Conf. Mach. Learn., ICML, Workshop Conf. Proc.*, Jun. vol. 32, 2014, pp. 1188–1196.

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[40] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. 30: Annu. Conf. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5998–6008.

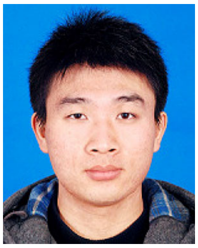
[41] S. Rendle, "Factorization machines with libFM," *ACM TIST*, vol. 3, no. 3, pp. 57:1–57: 22, 2012.

[42] X. He *et al.*, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 173–182.

[43] X. Wang *et al.*, "Heterogeneous graph attention network," in *Proc. World Wide Web Conf.*, May 2019, pp. 2022–2032.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations, Conf. Track Proc.*, San Diego, CA, USA, May 2015, pp. 1–15.

Desheng Cai received the bachelor's degree in computer science and technology from the Anhui University of Science and Technology, Huainan, China, in 2015 and the master's degree in computer science from Hefei University of Technology, Hefei, China, where he is currently working toward the Ph.D. degree. He is also with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include data mining, Recommendation, and Web intelligence.



Shengsheng Qian received the B.E. degree from Jilin University, Changchun, China, in 2012 and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2017. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His current research interests include social media data mining and social event content analysis.

Quan Fang received the B.E. degree from Beihang University, Beijing, China, in 2010 and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an Associate Professor with the Institute of Automation, Chinese Academy of Sciences. His research interests include georeferenced social media mining and application, multimedia content analysis, knowledge mining, computer vision, and pattern recognition.

Changsheng Xu (Fellow, IEEE) is a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China and the Executive Director of the China-Singapore Institute of Digital Media, Singapore. He has hold 30 granted or pending patents and authored or coauthored more than 200 refereed research papers in his areas of research, which include multimedia content analysis, indexing, and retrieval, pattern recognition, and computer vision. He is an Associate Editor for the IEEE Transaction on Multimedia, *ACM Transaction on Multimedia Computing, Communications and Applications,* and *ACM/Springer Multimedia Systems Journal*. He was the recipient of the Best Associate Editor Award of ACM Transaction on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He was the Program Chair of ACM Multimedia 2009. He was an Associate Editor, the Guest Editor, the General Chair, the Program Chair, the Area or Track Chair, the special session Organizer, the Session Chair, and a TPC Member of more than 20 IEEE and ACM prestigious multimedia journals, conferences and workshops. He is IAPR Fellow and ACM Distinguished Scientist.