What Aspect Do You Like: Multi-scale Time-aware User Interest Modeling for Micro-video Recommendation

Hao Jiang
Shandong University
jianghaosdu@gmail.com

Wenjie Wang*
National University of Singapore
wenjiewang96@gmail.com

Yinwei Wei
Shandong University
weiyinwei@hotmail.com

Zan Gao
Shandong Artificial Intelligence Institute
gaozan114@126.com

Yinglong Wang
Shandong Artificial Intelligence Institute
wangyl@sdas.org

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

ABSTRACT

Online micro-video recommender systems aim to address the information explosion of micro-videos and make the personalized recommendation for users. However, the existing methods still have some limitations in learning representative user interests, since the multi-scale time effects, user interest group modeling, and false positive interactions are not taken into consideration. In view of this, we propose an end-to-end Multi-scale Time-aware user Interest modeling Network (MTIN). In particular, we first present an interest group routing algorithm to generate fine-grained user interest groups based on user's interaction sequence. Afterwards, to explore multi-scale time effects on user interests, we design a time-aware mask network and distill multiple temporal information by several parallel temporal masks. And then an interest mask network is introduced to aggregate fine-grained interest groups and generate the final user interest representation. At last, in the prediction unit, the user representation and micro-video candidates are fed into a deep neural network (DNN) for predictions. To demonstrate the effectiveness of our method, we conduct experiments on two publicly available datasets, and the experimental results demonstrate that our proposed model achieves substantial gains over the state-of-the-art methods.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Personalization**; **Video search**; *Multimedia information systems.*

KEYWORDS

Micro-video Recommendation; User Interest Modeling; Temporal Attention Network

ACM Reference Format:
Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What Aspect Do You Like: Multi-scale Time-aware User Interest Modeling for Micro-video Recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413653

1 INTRODUCTION

In recent years, the recommender system has played an increasingly important role in micro-video sharing platforms, such as Tiktok, Kuai, and Instagram. This is attributed to the powerful ability of the recommender system to reduce user retrieval time and alleviate information overload. To pursue a high-quality micro-video recommender, it is crucial to discover user interests and provide micro-videos according to their tastes. A statistical report from Jiguang company[1] shows that 73.9% of 113 million platform users express as many as 20 types of interests. As such, in order to provide better user experience, it is essential to further explore their personalized interests.

With the rapid development of deep learning technology, many efforts have been devoted to obtaining personalized user interests in the field of recommendation [12, 17, 19, 21, 35, 36, 38, 42, 43], especially the combination of recurrent neural networks (RNN) [8, 22] and attention mechanisms [4, 30], which has made great progress in capturing long-term and short-term user preferences and learning diverse interests. For example, Chen *et al.* [2] proposed a hierarchical attention network at category-level and item-level for long-term and short-term interest modeling in micro-video click-through prediction. Li *et al.* [19] utilized a multi-interest extractor layer to capture diverse user interests and built a label-aware attention layer for personalized recommendation.

Despite their remarkable performance, there are still some issues untouched, which are summarized as follows:

- **Multi-scale time effects.** Previous methods usually consider that the effect of micro-videos on user interest modeling decreases over time implicitly, which is captured by RNN [8, 10, 11] or learned from timestamp features [20, 30]. However, they ignore the case that the importance of micro-videos decreases over time varies from user to user, that is to say, for different users,

* Wenjie Wang and Liqiang Nie are corresponding authors.

[1] https://www.jiguang.cn/reports/43.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413653

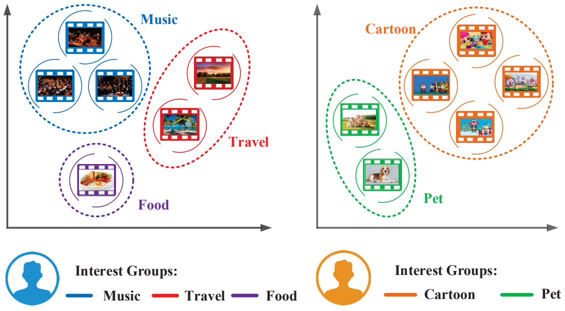


Figure 1: An illustration of personalized interest groups for different users.

time has various effects on their interests. Therefore, it is necessary to explicitly model multi-scale time effects on user interest modeling for micro-video recommendation.

- **User interest group modeling.** As shown in Figure 1, different users have different interest groups and the interest groups of one user may also be diverse. However, existing methods commonly focus on learning the user interest representation directly from micro-video features without grouping [4, 8, 15, 25]. In this way, the large groups with a majority of historical micro-videos dominate the user interests, and the micro-videos in small groups are rarely recommended. Inspired by this, we argue that fine-grained grouping for micro-videos is important to capture diverse user interests.
- **False-positive interactions.** Considering that the explicit feedback (*e.g.*, rating, review) is not always available, implicit feedback, such as clicks and browsing, tends to be used to train the recommendation model [2, 33, 34, 38]. However, for implicit feedback, some micro-videos clicked by the user may not indicate the user's real interests, which is named as *false-positive interactions*. For example, a user watches a micro-video shared by her/his friends, while she/he has no interest in this micro-video. We argue that the false-positive interactions harm the user interest modeling, while how to deal with the detrimental effects has not been considered in previous methods.

Indeed, it is tough to address the aforementioned problems due to the following challenges: 1) Given that it is already non-trivial to model the temporal information in explicit ways, explicitly capturing the multi-scale time effects in user interest modeling becomes more difficult. 2) Each user has her/his personalized interests, so it is non-trivial to learn fine-grained interests of different users and distill their interest groups. And 3) different from explicit feedback, implicit feedback has noises [13] and does not always indicate the actual user interests, which poses a huge challenge in distinguishing false-positive interactions and clean interactions.

To overcome these challenges, we present an end-to-end Multi-scale Time-aware user Interest modeling Network (MTIN), as shown in Figure 2. It consists of three units — the interest group routing unit, the item-level and group-level interest extraction unit, and the prediction unit. Specifically, 1) in the first unit, we propose an interest group routing algorithm, which is used to generate user interest groups based on the interaction sequence. Meanwhile, we introduce a discount factor to reduce the detrimental effects of false-positive interactions. 2) In the second unit, to explore multi-scale time effects on user interests, we design a temporal mask network which takes the results of the interest group routing unit as input and distills multi-scale time effects by several parallel temporal masks. The output group representations are adopted to the interest mask network, which is used to aggregate fine-grained interest groups and generate the final user interest representation. And 3) in the third unit, the user interest representation and micro-video candidates are fed into a deep neural network (DNN) for predictions. To demonstrate our proposed model, we conduct extensive experiments on two publicly available datasets. The results show that our proposed model outperforms several state-of-the-art baselines.

The main contributions are summarized as follows:

- To explicitly exploit the multi-scale time effects in user interest modeling, we develop a parallel temporal mask network, which is able to learn multiple temporal information for micro-video recommendation.
- To learn diverse user interests, we propose an interest group routing algorithm, which is capable of generating fine-grained user interest groups. In addition, we introduce an interest mask network to aggregate interest groups and distill the final user interest representation.
- We conduct extensive experiments on two publicly available datasets, verifying the effectiveness of our proposed MTIN over the state-of-the-art methods. In addition, we release our codes and involved parameters[2] to benefit other researchers.

2 RELATED WORK

In this section, we review the methods related to our research, including video recommendation and user interest modeling.

2.1 Video Recommendation

The existing methods of video recommendation can be divided into three categories: collaborative filtering methods [9, 31, 32], content-based methods [3, 5, 6, 33, 37] and hybrid approaches [1, 41]. Collaborative filtering (CF) is widely used in recommender systems, which models user interests by exploring user-item interactions with the assumption that people with similar interests tend to make similar choices. For example, Huang *et al.* [15] developed a real-time matrix factorization based CF algorithm with an adjustable online updating strategy for video recommendation. However, CF methods suffer from the problem of cold start and data sparsity [7, 25, 31]. To tackle this issue, some researchers adopted the methods of integrating user/item content representations into feature extractions [23, 24, 26, 27], *i.e.*, content-based filtering. For example, Cui *et al.* [6] studied the video representations with social attributes and users with content attributes by harvesting video propagation traces among users-item interactions. As for hybrid approaches, researchers combined both CF and content-based methods for recommendation. For instance, Chen *et al.* [1] employed an attention mechanism in CF to address the item-level and component-level implicit feedback in multimedia recommendation. Despite their remarkable performance, they usually directly learn user interests from item features without grouping, limiting the exploration of fine-grained user interests. Differently, our model explores user

[2] https://anonymous-1429.wixsite.com/anonymous.

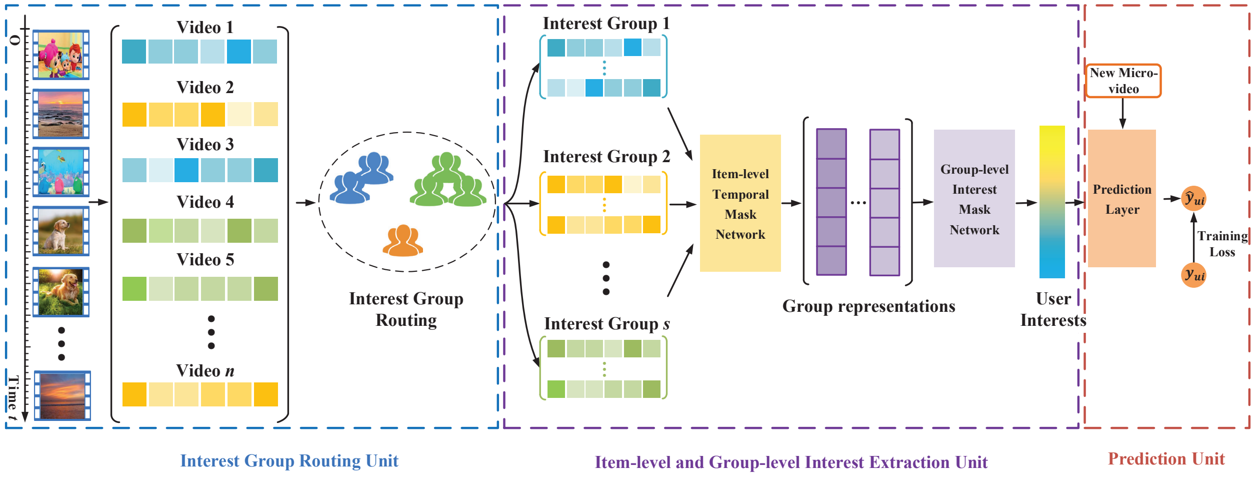


Figure 2: Illustration of our proposed multi-scale time-aware user interest modeling network for micro-video recommendation. It consists of three units: the interest group routing unit, the item-level and group-level interest extraction unit, and the prediction unit.

interests via an interest group routing algorithm to capture fine-grained user interest groups.

2.2 User Interest Modeling

User interest modeling is a common issue in recommender systems. Existing methods learn user interest representations from user profiles [5, 34], social relationships among users [16, 29], and user-item interactions [14, 39]. For example, Covington *et al.* [5] concatenated the discretized attributes (*e.g.*, age, gender) as original inputs of the neural network for video recommendation. Tran *et al.* [29] explored user interests by exploring user-user interactions based on sub-attention network for group recommendation. Zhou *et al.* [43] designed an attention network to obtain varying representations of user interests which depends on different target items. However, these methods limit to explore multi-scale time effects of user interests, while ignoring the detrimental effects of false-positive interactions. In light of this, we utilize the multi-scale temporal masks to explore multiple time effects, and introduce a discount factor in the interest group routing algorithm to deal with false-positive interactions.

3 PROBLEM DEFINITION

A micro-video recommender system contains two sets of entities: users and micro-videos. Let $\mathcal{U} = \{u_1, u_2, u_3, ..., u_{|\mathcal{U}|}\}$ denote the user set and $\mathcal{I} = \{i_1, i_2, i_3, ..., i_{|\mathcal{I}|}\}$ denote the micro-video set, where $|\mathcal{U}|$ and $|\mathcal{I}|$ denote the number of users in set $\mathcal{U}$ and items in set $\mathcal{I}$, respectively. Each interaction between the user and the micro-video is associated with a timestamp, which can be formulated as a triplet $i_t^{(u)} = \langle u, i, t \rangle$, where $u \in \mathcal{U}$, $i \in \mathcal{I}$, and t represents the timestamp when the interaction happens. By sorting the interaction records in ascending order according to the timestamp, an interaction sequence for user u can be represented as $\mathcal{H}_u = \{i_{t_1}^{(u)}, ..., i_{t_j}^{(u)}, ..., i_{t_L}^{(u)}\}$, where $i_{t_j}^{(u)} \in \mathcal{I}$ is the micro-video interacted by user u at time t_j and L is the length of the interaction sequence. Furthermore, the interaction sequence $\mathcal{H}_u$ is divided into $\mathcal{H}_{pos}$ and $\mathcal{H}_{neg}$, which respectively represent the micro-videos clicked by the user and the ones whose thumbnails browsed by the user but not clicked.

Formally, the micro-video recommendation task can be defined as follows:

Input: The user set $\mathcal{U}$, the interaction sequence $\mathcal{H}_u$ of each user u, and the candidate micro-video $i_{t_{L+1}}^{(u)}$.

Output: The probability $Prob\left(i_{t_{L+1}}^{(u)}\right)$ that the new micro-video will be clicked by user u, which is formulated as:

$$Prob\left(i_{t_{L+1}}^{(u)}\right) = \mathcal{F}(u, \mathcal{H}_u, i_{t_{L+1}}^{(u)}), \tag{1}$$

where $\mathcal{F} : \mathcal{I} \mapsto \mathbb{R}$ denotes the prediction function.

4 OUR PROPOSED MODEL

In this section, we introduce the architecture of our proposed model. As shown in Figure 2, our model consists of three units — the interest group routing unit, the item-level and group-level interest extraction unit, and the prediction unit. To be more specific, the first unit is designed for generating user interest groups based on the interaction sequence. The second unit is proposed to leverage fine-grained interest groups and distill user interest representations. And the third unit aims to predict the click probabilities of micro-video candidates. In the following parts, we describe each unit in detail.

4.1 Interest Group Routing Unit

The interest group routing process in this unit consists of two steps: 1) calculating item-group matching scores, and 2) assigning interest groups for micro-videos.

4.1.1 *Calculating Item-Group Matching Scores.*

Based on the positive historical interaction sequence $\mathcal{H}_{pos}$ of user u, we represent

Algorithm 1 Interest Group Routing Algorithm.

```
Input:
    User's positive interaction sequence $\mathcal{H}_{pos}$;
    Matching scores $\mathcal{P} = \bigcup_{j=1}^{l_0} \mathcal{P}_j = \{p_{1\leftarrow j}, p_{2\leftarrow j}, ..., p_{s\leftarrow j}\}$;
    Iteration number $\tau$;
Output:
    Interest groups $\mathcal{E} = \bigcup_{g=1}^{s} \mathcal{E}_g = \bigcup_{g=1}^{s} \{i_1^{(g)}, i_2^{(g)}, ..., i_l^{(g)}\}$;
1:  for each $i_j \in \mathcal{H}_{pos}$ do
2:      $S_j = -1, \mathcal{E}_{S_j} \leftarrow \emptyset$;
3:  end for
4:  for each iteration do
5:      for each $i_j \in \mathcal{H}_{pos}$ do
6:          $\xi_g = \log_b(b + \max(\mathrm{avg}_{\mathcal{E}_g} - p_{g\leftarrow j}, 0)), \quad g \in [1, s]$;
7:          $p_{g\leftarrow j}^{(d)} = p_{g\leftarrow j}/\xi_g, \quad g^* = \underset{g}{\mathrm{argmax}}(p_{g\leftarrow j}^{(d)})$;
8:          if $g^* \neq S_j \wedge p_{g\leftarrow j}^{(d)} > \epsilon$ then
9:              $\mathcal{E}_{S_j} \leftarrow \mathrm{POP}(\mathcal{E}_{S_j}, i_j), \quad S_j = g^*, \quad \mathcal{E}_{g^*} \leftarrow \mathrm{ADD}(\mathcal{E}_{g^*}, i_j)$
10:         end if
11:     end for
12: end for
13: $\mathcal{E}_g \xleftarrow{\text{SortByTime}} \mathcal{E}_g = \{i_1^{(g)}, i_2^{(g)}, ..., i_l^{(g)}\}, \quad g \in [1, s]$;
14: return $\mathcal{E} = \bigcup_{g=1}^{s} \mathcal{E}_g$
```

each micro-video j in $\mathcal{H}_{pos}$ as an embedding vector $\boldsymbol{x}_j \in \mathbb{R}^d$, where d is the embedding size. And we pretrain a positive interest memory matrix $\boldsymbol{M}_u \in \mathbb{R}^{s\times d}$ for user u, where s denotes the number of interest groups. Besides, we employ $\boldsymbol{c}_g \in \mathbb{R}^d$ to represent the interest group embedding in $\boldsymbol{M}_u$, *i.e.*, $\boldsymbol{M}_u = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_g, ..., \boldsymbol{c}_s]^T$.

In order to calculate item-group matching scores, we utilize an item-group co-attention network. Firstly, given the micro-video embedding $\boldsymbol{x}_j$ and the group embedding $\boldsymbol{c}_g$, we calculate the co-attention matrix $\mathcal{S} \in \mathbb{R}^{l_0\times s\times d}$ between the micro-video and the interest group, where $l_0 = |\mathcal{H}_{pos}|$ denotes the length of user's positive interaction sequence. Specifically, the entry $\boldsymbol{s}_{j,g}$ of matrix $\mathcal{S}$ is calculated by:

$$\boldsymbol{s}_{j,g} = \boldsymbol{W}_s\sigma(\boldsymbol{W}_v\boldsymbol{x}_j + \boldsymbol{b}_v + \boldsymbol{W}_u\boldsymbol{c}_g + \boldsymbol{b}_u) + \boldsymbol{b}_s, \tag{2}$$

where $\boldsymbol{W}_s, \boldsymbol{W}_v, \boldsymbol{W}_u \in \mathbb{R}^{d\times d}$ are weight matrices, $\boldsymbol{b}_v, \boldsymbol{b}_u, \boldsymbol{b}_s \in \mathbb{R}^d$ are biases, and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is the element-wise activation function. Based on this, the micro-video embedding $\boldsymbol{x}_j$ and group embedding $\boldsymbol{c}_g$ are weighted by $\boldsymbol{s}_{j,g}$:

$$\begin{cases} \widehat{\boldsymbol{x}_j} = \boldsymbol{s}_{j,g} \odot \boldsymbol{x}_j, \\ \widehat{\boldsymbol{c}_g} = \boldsymbol{s}_{j,g} \odot \boldsymbol{c}_g, \end{cases} \tag{3}$$

where $\odot$ is the element-wise product operation. And we get the matching score $p_{g\leftarrow j}$ between micro-video j and group g by:

$$p_{g\leftarrow j} = \boldsymbol{W}_p^h\sigma(\boldsymbol{W}_p^l[\widehat{\boldsymbol{x}_j}, \widehat{\boldsymbol{c}_g}] + \boldsymbol{b}_p^l) + b_p^h, \tag{4}$$

where $\boldsymbol{W}_p^h \in \mathbb{R}^d$, $\boldsymbol{W}_p^l \in \mathbb{R}^{d\times 2d}$ are weight matrices, $\boldsymbol{b}_p^l \in \mathbb{R}^d$, $b_p^h \in \mathbb{R}$ are biases, $[\cdot]$ is the concatenation operation, and $\sigma(\cdot)$ is the element-wise activation function. The matching score $p_{g\leftarrow j}$ is one of the inputs of our interest group routing algorithm, which is used for the interest group assignments of the micro-video sequence.

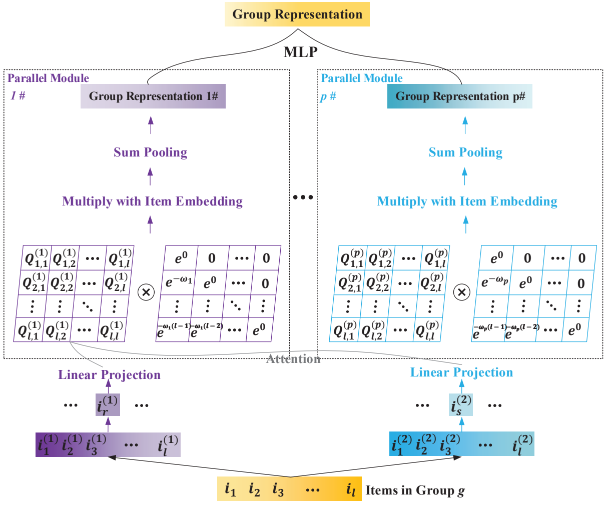

Group Representation
MLP
Parallel Module
1 #
Group Representation 1#
Sum Pooling
Multiply with Item Embedding
$Q_{1,1}^{(1)}$ $Q_{1,2}^{(1)}$ ⋯ $Q_{1,l}^{(1)}$
$Q_{2,1}^{(1)}$ $Q_{2,2}^{(1)}$ ⋯ $Q_{2,l}^{(1)}$
$Q_{l,1}^{(1)}$ $Q_{l,2}^{(1)}$ ⋯ $Q_{l,l}^{(1)}$
⊗
e^0 0 ⋯ 0
$e^{-\omega_1}$ e^0 ⋯ 0
$e^{-\omega_1(l-1)}$ $e^{-\omega_1(l-2)}$ ⋯ e^0
Parallel Module
p #
Group Representation p#
Sum Pooling
Multiply with Item Embedding
$Q_{1,1}^{(p)}$ $Q_{1,2}^{(p)}$ ⋯ $Q_{1,l}^{(p)}$
$Q_{2,1}^{(p)}$ $Q_{2,2}^{(p)}$ ⋯ $Q_{2,l}^{(p)}$
$Q_{l,1}^{(p)}$ $Q_{l,2}^{(p)}$ ⋯ $Q_{l,l}^{(p)}$
⊗
e^0 0 ⋯ 0
$e^{-\omega_p}$ e^0 ⋯ 0
$e^{-\omega_p(l-1)}$ $e^{-\omega_p(l-2)}$ ⋯ e^0
Attention
Linear Projection
$i_r^{(1)}$
$i_1^{(1)}$ $i_2^{(1)}$ $i_3^{(1)}$ ⋯ $i_l^{(1)}$
Linear Projection
$i_s^{(2)}$
$i_1^{(2)}$ $i_2^{(2)}$ $i_3^{(2)}$ ⋯ $i_l^{(2)}$
i_1 i_2 i_3 ⋯ i_l Items in Group g

Figure 3: Illustration of the item-level temporal mask network, which is used for exploring multi-scale time effects on user interests and generating interest group representations.

4.1.2 *Assigning Interest Groups.* To assign interest groups based on user's interaction sequence, we propose an interest group routing algorithm, which takes the matching scores between micro-videos and groups as input and outputs the user interest group assignment results. The process is detailed in Algorithm 1.

At first, we use the interest group pointer S_j to represent the assigned group for each micro-video j in $\mathcal{H}_{pos}$, and initialize all pointers S_j and group sets $\mathcal{E}_{S_j}$ before iterations (Line 2). As mentioned earlier, some false-positive interactions are mixed in the user's interaction sequence and interfere with user interest modeling. To address this problem, we introduce a discount factor ξ_g to filter out the false-positive interactions (Line 6), where $avg_{\mathcal{E}_g}$ is the average score of micro-videos in group g. In this way, we recalculate matching scores $p_{g\leftarrow j}^{(d)}$ based on $p_{g\leftarrow j}$ and ξ_g, and then select the interest group g^* with the highest matching score $p_{g\leftarrow j}^{(d)}$ as the assignment target of the micro-video j (Line 7). Afterwards, we compare $p_{g\leftarrow j}^{(d)}$ with ϵ to make the next decision of interest group assignment, where ϵ is a manually adjusted threshold (Line 8). If the condition in Line 8 is true, we pop the micro-video j from the previous interest group set $\mathcal{E}_{S_j}$ and assign it to the new interest group set $\mathcal{E}_{g^*}$ (Line 9). After the iterations, we sort micro-videos in each group according to their interaction timestamps (Line 13), and return $\mathcal{E} = \bigcup_{g=1}^{s} \mathcal{E}_g$ as the obtained interest groups (Line 14).

On the one hand, the proposed algorithm has the ability to generate fine-grained interest groups based on user's interaction sequence. On the other hand, it handles false-positive interactions. It is easy to understand that when we try to assign groups for the micro-video that does not belong to the user interests, the discount factor will largely decrease the matching scores, preventing it from being assigned to any interest group and reducing its potential detrimental effects on user interest modeling.

In fact, some micro-video platforms provide category labels for micro-videos. The reason why we do not directly adopt the categories is that they are not always available (such as micro-videos uploaded by users without any category information). In addition, extensive categories may produce many interest groups, while the number of groups in the group routing algorithm is controllable. The effective way to combine the category information and grouping algorithm could be explored in future work.

4.2 Item-level Temporal Mask Network

In order to capture multi-scale time effects and get representations of each interest group, we introduce an item-level temporal mask network, as shown in Figure 3, which learns group representations based on micro-video features. It takes the obtained groups of micro-videos from the interest group routing unit (*i.e.*, $\mathcal{E} = \bigcup_{g=1}^{s} \mathcal{E}_g = \bigcup_{g=1}^{s} \{i_1^{(g)}, i_2^{(g)}, ..., i_l^{(g)}\}$) as input, and outputs the interest group representations.

Many previous studies [2, 20, 40] have made a lot of efforts in modeling the sequential information of user's interaction sequence for recommendation. However, for different users, time has various effects on their interests. Previous studies ignore the case that the importance of micro-videos decreases over time changes from user to user, and we argue that the multi-scale time effects in user interest modeling have not been explicitly considered. To address this problem, we design an item-level temporal mask network to explore multi-scale time effects on user interests. Specifically, the parallel temporal masks are utilized to capture multiple temporal information of user's historical interactions.

Formally, we denote the embedding vector of the micro-video in interest groups as $\bar{\mathcal{E}} = \bigcup_{g=1}^{s} \bar{\mathcal{E}}_g = \bigcup_{g=1}^{s} \{\boldsymbol{x}_1^{(g)}, \boldsymbol{x}_2^{(g)}, ..., \boldsymbol{x}_l^{(g)}\}$. In the k-th parallel temporal mask, we first calculate the attention score $Q_{i,j}^{(k)}$ among the micro-video pairs in group g based upon the following formula[3]:

$$Q_{i,j}^{(k)} = (\boldsymbol{x}_i)^T \boldsymbol{W}_g^k \boldsymbol{x}_j, \tag{5}$$

where k and $\boldsymbol{W}_g^k \in \mathbb{R}^{d \times d}$ denote the identifier of the k-th parallel temporal mask and the weight matrix, respectively.

For the k-th parallel temporal mask, the element at each position in the mask is calculated by the following formula:

$$P_{i,j}^{(k)} = \begin{cases} e^{-(i-j)\omega_k} & i > j \\ 0 & i \leq j, \end{cases} \tag{6}$$

where i, j are identifiers of the row and the column in the temporal mask, respectively. ω_k denotes the coefficient in the k-th temporal mask.

We design the temporal masks due to the following reasons: 1) Motivated by [21], we choose an exponential form for the temporal mask, which is able to describe the gradual decay of the importance of past interactions as time goes. And 2) the exponential function in the temporal mask can fit multiple importance decreases of historical micro-videos with different parameters, and the parallel temporal masks are able to explore multi-scale time effects on user interest modeling.

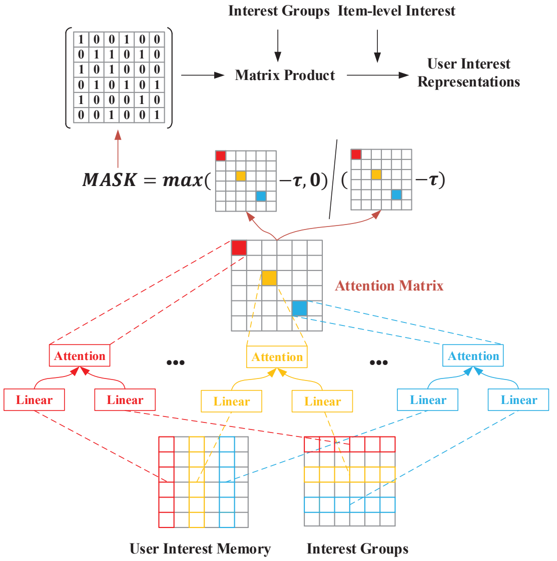


Figure 4: Illustration of the group-level interest mask network, which is used for aggregating fine-grained interest groups and distilling the final user interest representations.

According to Formula 7, we obtain the interest group representation based on micro-video features. In the temporal mask, we use the weighted sum of historical micro-video features to obtain the current micro-video representation. Afterwards, each parallel module aggregates all micro-video features in the group through a sum pooling operation. We concatenate the output of each parallel module, and then generate the group representation by a multi-layer perception (MLP) network:

$$\begin{cases} M_{i,j}^k = P_{i,j}^{(k)} Q_{i,j}^{(k)}, \\ \boldsymbol{m}_i^k = \sum_{j=1}^{l} M_{i,j}^k \boldsymbol{x}_j, \\ \boldsymbol{r}_g^k = \sum_{i=1}^{l} \boldsymbol{m}_i^k, \\ \boldsymbol{r}_g = MLP([\boldsymbol{r}_g^1, \boldsymbol{r}_g^2, ..., \boldsymbol{r}_g^k, ..., \boldsymbol{r}_g^p]), \end{cases} \tag{7}$$

where $\boldsymbol{m}_i^k$ and $\boldsymbol{r}_g^k$ denote the representation of micro-video i and interest group g generated by k-th parallel temporal mask, respectively. p is the number of parallel modules, and $\boldsymbol{r}_g$ denotes the final group representation. In this way, we get the representation of each interest group.

4.3 Group-level Interest Mask Network

Using the generated interest group representation, we next understand user interest based on our proposed group-level interest mask network. As illustrated in Figure 4, the network is used for aggregating fine-grained interest groups and distilling the final user interest representations. Specifically, it takes the group representation $\boldsymbol{r}_g$ as input, and outputs the user interest representation.

The design of this network has the following considerations: 1) user interests can be divided into different interest groups. These fine-grained group representations are combined with different

[3] For simplicity, the interest group identifier is omitted here.

weights to form user interest representations. 2) The information in an interest group is always complicated and the implicit feedback has noises [13]. As such, perhaps not all of the information in the group is closely related to the interests of users. In order to avoid poor recommendations caused by the useless information, we introduce an interest mask to filter it out.

We have obtained the interest memory matrix $\boldsymbol{M}_u = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_p, ..., \boldsymbol{c}_s]^T$ (Section 4.1.1) and the group representations $\mathcal{R} = \{\boldsymbol{r}_1, ..., \boldsymbol{r}_g, ..., \boldsymbol{r}_s\}$ (Section 4.2). We calculate the attention score $w_{p,g}$ between memory vector $\boldsymbol{c}_p$ and group representation $\boldsymbol{r}_g$ as follows:

$$w_{p,g} = \frac{\exp(\boldsymbol{c}_p^T \boldsymbol{W}_h \boldsymbol{r}_g)}{\sum_{g^*=1}^{s} \exp(\boldsymbol{c}_p^T \boldsymbol{W}_h \boldsymbol{r}_{g^*})}, \tag{8}$$

where $\boldsymbol{W}_h$ is the trainable matrix. Next we introduce the group-level filter factor $\hat{w}_{p,g}$, which is defined as follows:

$$\hat{w}_{p,g} = \frac{max(w_{p,g} - \tau, 0)}{w_{p,g} - \tau}. \tag{9}$$

where τ is a manually adjusted hyper parameter. From Formula 9, we get an interest mask consisting of $\hat{w}_{p,g}$ with elements 0 and 1. If $w_{p,g}$ is greater than τ, $\hat{w}_{p,g}$ will be set to 1, otherwise 0. After getting $\hat{w}_{p,g}$, the group-level user interest representations of p-th interest group are calculated as follows:

$$\boldsymbol{h}_p = \sum_{g=1}^{s} \hat{w}_{p,g} \boldsymbol{r}_g. \tag{10}$$

At last, the item-level user interest (*i.e.*, $\boldsymbol{r}_p$) and the group-level user interest (*i.e.*, $\boldsymbol{h}_p$) are aggregated (*e.g.*, sum pooling) to obtain user representations of p-th interest group (*i.e.*, $\boldsymbol{c}_p$), which are eventually added to the corresponding interest group in $\boldsymbol{M}_u$.

Note that the above method treats the positive interaction in $\mathcal{H}_{pos}$ and the negative interaction in $\mathcal{H}_{neg}$ separately. In the above discussion, we take the positive interaction as an example.

4.4 Model Prediction and Optimization

Given the user's interest memory matrix $\boldsymbol{M}_u$ and the new micro-video's embedding vector $\boldsymbol{x}_e$, we aggregate the memory vectors of different interest groups in $\boldsymbol{M}_u$ to get the user's positive interest representation $\boldsymbol{q}$ by the sum pooling operation:

$$\boldsymbol{q} = sumpooling(\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_s). \tag{11}$$

We concatenate the user vector $\boldsymbol{q}$ and the item vector $\boldsymbol{x}_e$ together, and then feed them into two MLP layers to calculate the prediction score $P(\boldsymbol{x}_e|\mathcal{H}_{pos})$ based on the positive interaction sequence.

In the same way as calculating $P(\boldsymbol{x}_e|\mathcal{H}_{pos})$, we calculate the prediction score $P(\boldsymbol{x}_e|\mathcal{H}_{neg})$ based on the negative interaction sequence [22], which aims to maximize the distance between the new micro-video embedding and user's negative interest features.

The final recommendation probability $\hat{y}_{ij}$ is represented by the linear combination of $P(\boldsymbol{x}_e|\mathcal{H}_{pos})$ and $P(\boldsymbol{x}_e|\mathcal{H}_{neg})$. And the objective function of our model is as follows:

$$\mathcal{L} = -\sum_{i \in \mathcal{U}} \left(\sum_{j \in \mathcal{H}_{pos}^i} \log\sigma(\hat{y}_{ij}) + \sum_{j \in \mathcal{H}_{neg}^i} \log(1 - \sigma(\hat{y}_{ij})) \right) + \lambda ||\Theta||_2^2, \tag{12}$$

Table 1: Statistics of the datasets.

Datasets	#Users	#Micro-videos	#Train Int.	#Test Int.	#Total Int.
MicroVideo-1.7M	10,986	1,704,880	8,970,310	3,767,309	12,737,619
KuaiShou-Dataset	10,000	3,239,534	10,931,092	2,730,291	13,661,383

where $\hat{y}_{ij}$ denotes the prediction score of micro-video j for user i, σ represents the sigmoid activation function, Θ refers to the set of parameters to be regularized, and λ is the regularization factor.

5 EXPERIMENTS

In this section, we conduct experiments on two publicly available datasets to evaluate the effectiveness of our proposed model. We aim to answer the following research questions:

- **RQ1:** How does MTIN perform compared with the state-of-the-art methods?
- **RQ2:** How do different feature aggregation methods affect the performance of MTIN?
- **RQ3:** How do the designed interest memory matrix and user interest group modeling affect the performance of MTIN?
- **RQ4:** Are the fine-grained user interests effective in MTIN?

5.1 Experimental Settings

Datasets. To evaluate the performance of our model, we conduct experiments on two publicly available datasets: MicroVideo-1.7M [2] and KuaiShou-Dataset [22]. In these two datasets, each micro-video is associated with its visual features, and each interaction consists of user ID, micro-video ID, and the relative timestamp. The interactions between users and micro-videos are divided into positive interactions (*i.e.*, the user clicks the micro-video) and negative interactions (*i.e.*, the user browses the thumbnail but does not click it). Dataset settings (*e.g.*, the division of the training set and the test set) are based on the mainstream settings [2, 22]. And the statistics of the datasets are summarized in Table 1.

Baselines. We consider the following methods for performance comparison:

- **BPR** [28]. Bayesian personalized ranking uses the pairwise ranking loss in the Bayesian approach to learn the relative ranking of positive and negative items of each user.
- **LSTM** [40]. Long short-term memory (LSTM) can be utilized to model the sequential information, and we use the aggregation of hidden states of each unit to form the user interest representations.
- **CNN**. We implement the convolutional neural network (CNN) to generate user interest representations based on the interaction sequence. The max pooling layer and MLP layers are used for user interest extraction and prediction.
- **NCF** [9]. NCF is used to model the interactions between latent features of users and items by replacing the inner product with the neural architecture, which is able to learn an arbitrary function from data.
- **ATRank** [42]. ATRank is an attention based recommendation framework, which projects all types of behavior into multiple latent semantic spaces via a self-attention mechanism to consider the heterogeneous user behaviors.

Table 2: Overall Performance Comparison. The best results are highlighted in bold, and significant improvements over the best baseline results are marked with † (t-test, p<0.01).

Methods	MicroVideo-1.7M				KuaiShou-Dataset			
	AUC	Precision@50	Recall@50	F1-score@50	AUC	Precision@50	Recall@50	F1-score@50
BPR	0.583	0.241	0.181	0.206	0.595	0.290	0.387	0.331
LSTM	0.641	0.277	0.205	0.236	0.713	0.316	0.420	0.360
CNN	0.650	0.287	0.214	0.245	0.719	0.312	0.413	0.356
NCF	0.672	0.316	0.225	0.262	0.724	0.320	0.420	0.364
ATRank	0.660	0.297	0.221	0.253	0.722	0.322	0.426	0.367
THACIL	0.684	**0.324**	0.234	0.269	0.727	0.325	0.429	0.369
ALPINE	0.713	0.300	0.460	0.362	0.739	0.331	0.436	0.376
MTIN	**0.729**†	0.317	**0.476**†	**0.381**†	**0.752**†	**0.341**†	**0.449**†	**0.388**†

- **THACIL** [2]. THACIL utilizes temporal windows and the multi-head self-attention to capture short-term and long-term user interests for micro-video click-through prediction.
- **ALPINE** [22]. ALPINE models user's dynamic and diverse interests by a temporal graph-guided approach. In addition, this method learns the enhanced representation of users by considering multi-level user interests.

Evaluation Metrics and Parameter Settings. To evaluate our proposed model, we adopt the evaluation metrics which are widely used in previous work [36, 42], including the Area Under Curve (AUC), Precision@K, Recall@K, and F1-measure@K. In our experiments, we set K = 50 and report the average scores on the test set. The user embedding and micro-video embedding are 128-dimensional vectors. We set the number of parallel modules to 8, the batch size to 32. The learning rate is set to 0.001, and the regularization factor is 0.0001. The number of interest groups is set to 4 on KuaiShou-Dataset and 6 on MicroVideo-1.7M. We optimize the parameters using Adam [18] optimizer.

5.2 Performance Comparison (RQ1)

The performance comparison of our model with the baselines on MicroVideo-1.7M and KuaiShou-Dataset is shown in Table 2. From the experimental results, we obtain the following observations:

- Our model achieves the best performance on the two datasets. Compared with ATRank, THACIL, and ALPINE, our model considers the interest group routing process in user interest modeling, and fine-grained user interest helps improve the performance. Moreover, by introducing parallel temporal masks, MTIN is capable of inferring multi-scale time effects on user interests, which is lacking in previous methods such as LSTM and CNN.
- BPR performs poorly on two datasets. LSTM consistently outperform BPR, demonstrating the importance of modeling the sequential information in user behaviors. NCF outperforms LSTM and CNN, which indicates the significance of nonlinear feature interactions between user embeddings and micro-video embeddings.
- User interest modeling methods (including ATRank, THACIL, and ALPINE) are superior to previous methods. Specifically, ATRank and THACIL show that the attention mechanism is beneficial for capturing user preference from historical interactions. And ALPINE performs better than THACIL, which demonstrates the necessity of modeling dynamic and diverse user interests. In addition, the modeling of multi-level user interests in ALPINE also helps to improve the model performance.

5.3 Study of Aggregation Methods (RQ2)

5.3.1 ***Effect of the Feature Aggregation Methods.*** We explore the effect of different feature aggregation methods of the item-level and group-level user interest representations. Specifically, we adopt three different feature aggregation methods, namely concatenation, sum pooling, and their combination. The experimental results of different feature aggregation methods on KuaiShou-Dataset are shown in Table 3. From the experimental results, we observe that the results of concatenation and sum pooling are lower than their combination. This indicates that the sum pooling and concatenation are not capable enough of exploring the competitive information interchange of item-level and group-level interest representations. And the combination of concatenation and sum pooling provides more powerful capabilities to capture high-order feature interactions and encode hidden relationships of user interests.

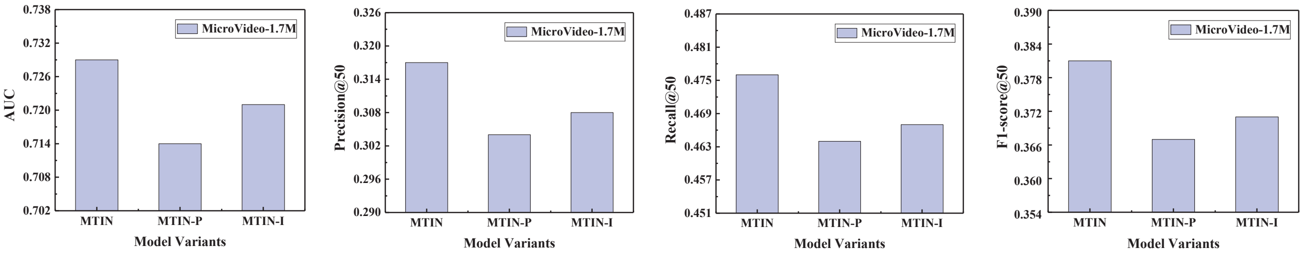
5.4 Study of MTIN (RQ3)

To evaluate the effectiveness of the interest memory matrix and interest group modeling of our proposed model, we conduct the ablation studies to compare MTIN with MTIN-P and MTIN-I. For the model variant MTIN-P, we remove the design of our pretrained interest memory matrix and only update it by group representations during the training process. And for the model variant MTIN-I, we remove the user interest group modeling and only make recommendations based on one positive interest group. We show the experimental results of MTIN, MTIN-P, and MTIN-I on the MicroVideo-1.7M dataset in Figure 5. According to the experimental results, we have the following observations:

- MTIN performs better than MTIN-P in terms of AUC, Precision, Recall, and F1-score, which demonstrates the effectiveness of our designed interest memory matrix. In our model, we first calculate the matching scores of micro-videos and

Table 3: Effect analysis of aggregation methods on KuaiShou-Dataset.

Aggregation Methods		TopK@10			TopK@50		
	AUC	Precision@10	Recall@10	F1-score@10	Precision@50	Recall@50	F1-score@50
concatenation	0.7520	0.3932	0.1108	0.1728	0.3403	0.4478	0.3867
sum pooling	0.7519	0.3915	0.1106	0.1725	0.3402	0.4473	0.3865
concatenation + pooling	**0.7524**	**0.3942**	**0.1111**	**0.1734**	**0.3414**	**0.4494**	**0.3880**



(a) AUC (b) Precision@50 (c) Recall@50 (d) F1-score@50

Figure 5: Illustration of the study of interest memory matrix and user interest group modeling of our model on MicroVideo-1.7M dataset.

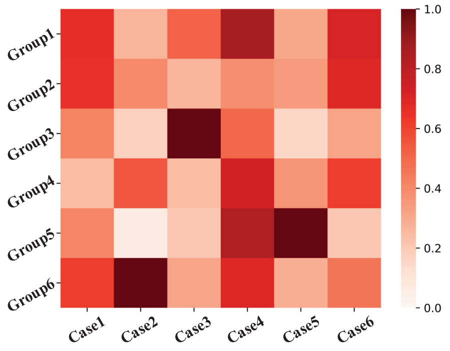


Figure 6: Heat map of the attention weights of each interest group on MicroVideo-1.7M dataset, which reflects the importance of different interest groups in expressing user interest.

interest groups based on the pretrained user interest memory matrix, and then learn the group representation of micro videos and update the user interest memory matrix, which makes MTIN achieve better performance.

- MTIN-I performs worse than MTIN, which indicates that exploring fine-grained interest groups is necessary in user interest modeling. From the experimental results of MTIN-I, we observe that only one positive user interest group is not sufficient to model various user interests, which makes MTIN-I perform poorly.

5.5 Study of Interest Groups (RQ4)

One of the considerations of our model is that users have personalized interest groups, and we aggregate fine-grained interest groups to extract user interest representations. In order to evaluate whether our model has the ability to learn fine-grained user interests, we visualized the relative importance of each group in user interest modeling on MicroVideo-1.7M dataset. Figure 6 presents the heat map of attention weights of each group corresponding to user interests, where different cases in the heat map represent different users in the dataset. From this heat map, we observe that in Case6, Group1, Group2 and Group4 have relatively dark colors, indicating that these three groups are more appealing to User6. This demonstrates that user interests are diverse and our model learns fine-grained user interest groups in user modeling. In addition, we observe that different users have different heat map distributions. For example, User2 is not interested in Group3, while User3 shows a strong interest in Group3, demonstrating that our model captures the personalized user interests to form user representation.

6 CONCLUSION AND FUTURE WORK

In this work, we focus on exploring the personalized user interests in micro-video recommendation. We propose MTIN, a multi-scale time-aware user interest modeling framework, which learns user interests from fine-grained interest groups. In particular, we incorporate the multi-scale time effects into user interests by time-aware parallel masks, and introduce the group routing algorithm to perform group assignments. Furthermore, extensive experiments on two publicly available datasets demonstrate that MTIN outperforms the state-of-the-art methods.

Our work provides some new ways for future research on micro-video recommender systems, such as exploring user social relationships for group routing and user interest learning, and incorporating micro-video semantics to understand user-item interactions. By integrating these considerations into our work, we could make a deeper understanding of user interests and establish more explainable and competitive recommender systems.

ACKNOWLEDGEMENTS

This work is supported by the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the Innovation Teams in Colleges and Universities in Jinan, No.:2018GXRC014; the National Natural Science Foundation of China, No.:U1936203.

REFERENCES

[1] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–344.

[2] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal hierarchical attention at category-and item-level for micro-video click-through prediction. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 1146–1153.

[3] Xusong Chen, Rui Zhao, Shengjie Ma, Dong Liu, and Zheng-Jun Zha. 2018. Content-Based Video Relevance Prediction with Second-Order Relevance and Attention Modeling. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 2018–2022.

[4] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. 2019. Co-Attentive Multi-Task Learning for Explainable Recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 2137–2143.

[5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191–198.

[6] Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What videos are similar with you? learning a common attributed representation for video recommendation. In *Proceedings of the 22nd ACM International Conference on Multimedia*. 597–606.

[7] Travis Ebesu, Bin Shen, and Yi Fang. 2018. Collaborative memory network for recommendation systems. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 515–524.

[8] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A unified personalized video recommendation via dynamic recurrent neural networks. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 127–135.

[9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*. ACM, 173–182.

[10] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *4th International Conference on Learning Representations*. OpenReview.net.

[11] Balázs Hidasi, Massimo Quadrana, Alexandros Karatzoglou, and Domonkos Tikk. 2016. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 241–248.

[12] Shintami Chusnul Hidayati, Cheng-Chun Hsu, Yu-Ting Chang, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. 2018. What dress fits me best? fashion recommendation on the clothing style for personal body shape. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 438–446.

[13] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *8th IEEE International Conference on Data Mining*. IEEE, 263–272.

[14] Xiaowen Huang, Quan Fang, Shengsheng Qian, Jitao Sang, Yan Li, and Changsheng Xu. 2019. Explainable Interaction-driven User Modeling over Knowledge Graph for Sequential Recommendation. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 548–556.

[15] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. 2016. Real-time video recommendation exploration. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, 35–46.

[16] Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the 4th ACM Conference on Recommender Systems*. ACM, 135–142.

[17] Hao Jiang, Wenjie Wang, Meng Liu, Liqiang Nie, Ling-Yu Duan, and Changsheng Xu. 2019. Market2Dish: A Health-aware Food Recommendation System. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2188–2190.

[18] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*. OpenReview.net.

[19] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 2615–2623.

[20] Chenliang Li, Xichuan Niu, Xiangyang Luo, Zhenzhong Chen, and Cong Quan. 2019. A review-driven neural model for sequential recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 2866–2872.

[21] Lei Li, Li Zheng, Fan Yang, and Tao Li. 2014. Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications* 41, 7 (2014), 3168–3177.

[22] Yongqi Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2019. Routing Micro-videos via A Temporal Graph-guided Recommendation System. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1464–1472.

[23] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards Micro-Video Understanding by Joint Sequential-Sparse Modeling. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 970–978.

[24] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28 (2019), 1235–1247.

[25] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-Video Co-Attention Network for Personalized Micro-video Recommendation. In *Proceedings of the 28th International Conference on World Wide Web*. ACM, 3020–3026.

[26] Liqiang Nie, Meng Liu, and Xuemeng Song. 2019. Multimodal learning toward micro-video understanding. *Synthesis Lectures on Image, Video, and Multimedia Processing* 9, 4 (2019), 1–186.

[27] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian. 2017. Enhancing micro-video understanding by harnessing external sounds. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 1192–1200.

[28] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 452–461.

[29] Lucas Vinh Tran, Tuan-Anh Nguyen Pham, Yi Tay, Yiding Liu, Gao Cong, and Xiaoli Li. 2019. Interact and decide: Medley of sub-attention networks for effective group recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 255–264.

[30] Huizhao Wang, Guanfeng Liu, An Liu, Zhixu Li, and Kai Zheng. 2019. DMRAN: A Hierarchical Fine-Grained Attention-Based Network for Recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 3698–3704.

[31] Meirui Wang, Pengjie Ren, Lei Mei, Zhumin Chen, Jun Ma, and Maarten de Rijke. 2019. A collaborative session-based recommendation approach with parallel memory modules. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 345–354.

[32] Pengfei Wang, Hanxiong Chen, Yadong Zhu, Huawei Shen, and Yongfeng Zhang. 2019. Unified Collaborative Filtering over Graph Embeddings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 155–164.

[33] Peng Wang, Yunsheng Jiang, Chunxu Xu, and Xiaohui Xie. 2019. Overview of Content-Based Click-Through Rate Prediction Challenge for Video Recommendation. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2593–2596.

[34] Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. 2020. Market2Dish: Health-aware Food Recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications* (2020).

[35] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2020. Denoising Implicit Feedback for Recommendation. *arXiv preprint arXiv:2006.04153* (2020).

[36] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized hashtag recommendation for micro-videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1446–1454.

[37] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. 2019. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing* 29 (2019), 1–14.

[38] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1437–1445.

[39] Xin Xin, Bo Chen, Xiangnan He, Dong Wang, Yue Ding, and Joemon Jose. 2019. CFM: Convolutional Factorization Machines for Context-Aware Recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. ijcai.org, 3926–3932.

[40] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential click prediction for sponsored search with recurrent neural networks. In *28th AAAI Conference on Artificial Intelligence*. AAAI Press.

[41] Xiaojian Zhao, Guangda Li, Meng Wang, Jin Yuan, Zheng-Jun Zha, Zhoujun Li, and Tat-Seng Chua. 2011. Integrating rich information for video recommendation with multi-task rank aggregation. In *Proceedings of the 19th ACM International Conference on Multimedia*. ACM, 1521–1524.

[42] Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiusi Chen, and Jun Gao. 2018. Atrank: An attention-based user behavior modeling framework for recommendation. In *32nd AAAI Conference on Artificial Intelligence*. AAAI Press.

[43] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 1059–1068.