Routing Micro-videos via A Temporal Graph-guided Recommendation System

Yongqi Li
Shandong University
liyongqi0@gmail.com

Meng Liu
Shandong University
mengliu.sdu@gmail.com

Jianhua Yin*
Shandong University
jhyin@sdu.edu.cn

Chaoran Cui
Shandong University of Finance and Economics
crcui@sdufe.edu

Xin-Shun Xu
Shandong University
xuxinshun@sdu.edu.cn

Liqiang Nie
Shandong University
nieliqiang@gmail.com

ABSTRACT

In the past few years, micro-videos have become the dominant trend in the social media era. Meanwhile, as the number of micro-videos increases, users are frequently overwhelmed by their uninterested ones. Despite the success of existing recommendation systems developed for various communities, they cannot be applied to routing micro-videos, since users in micro-video platforms have their unique characteristics: diverse and dynamic interest, multi-level interest, as well as true negative samples. To address these problems, we present a temporal graph-guided recommendation system. In particular, we first design a novel graph-based sequential network to simultaneously model users' dynamic and diverse interest. Similarly, uninterested information can be captured from users' true negative samples. Beyond that, we introduce users' multi-level interest into our recommendation model via a user matrix that is able to learn the enhanced representation of users' interest. Finally, the system can make accurate recommendation by considering the above characteristics. Experimental results on two public datasets verify the effectiveness of our proposed model.

CCS CONCEPTS

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

KEYWORDS

Micro-video Routing; Graph-based LSTM; Multi-level Interest; Negative Samples Modeling

ACM Reference Format:
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 (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350950

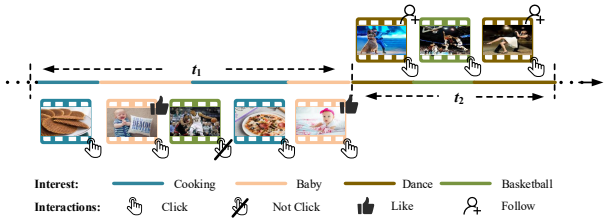


Figure 1: Illustration of a user's historical interactions with micro-videos, which reflects the user's diverse, dynamic, and multi-level interest.

1 INTRODUCTION

Owing to the proliferation of micro-video platforms, such as Instagram[1], Kuaishou[2], and Tik Tok[3], the amount of micro-videos generated and shared by people are growing exponentially. Considering Kuaishou as an example, as of January 2019, it had reached 190 million active users and 10 million uploaded micro-videos daily, and each online user usually spends up to one hour to share, search, and view their interested micro-videos[4]. Nevertheless, as the micro-videos surge, it becomes increasingly difficult and expensive for users to locate their desired micro-videos from the vast candidates. In the light of this, it is crucial to build a personalized recommendation system to intelligently route micro-videos to the target users.

Building a personalized recommendation system for micro-video services is non-trivial, due to the following reasons: 1) **Diverse and dynamic interest**. On the one hand, users' interest evolves over time, and it is hence a sequential expression. For example, as shown in Figure 1, a user likes cooking videos at time t_1, but may prefer dance videos at t_2. On the other hand, users' interest is diverse, namely a user may be fond of multiple topics at the same time. In a sense, personalized recommendation requires to simultaneously model users' dynamic and diverse interest information. 2) **Multi-level interest**. Users may have different interaction types on micro-videos, including "click", "like", and "follow", which signal different degrees of interest. For example, "click" means the user is attracted to the micro-video, "like" is one much enjoys and

*Corresponding author: Jianhua Yin.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350950

[1]https://www.instagram.com/.
[2]https://www.kuaishou.com/.
[3]https://www.tiktok.com/.
[4]http://www.sohu.com/a/295239939_441449.

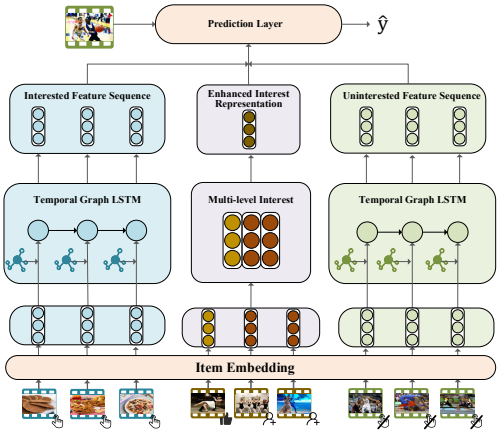


Figure 2: Illustration of our proposed ALPINE model.

appreciates the micro-video, and "follow" refers to the user likes the micro-video very much and wishes to see it again in future. Heretofore, how to integrate the various degrees of interest into personalized recommendation is largely untapped. And 3) **true negative samples**. As we know, prior methods commonly assume that nonpositive items are negative samples, which is hardly reliable to infer which item a user did not like. Different from these models, we are able to obtain true negative samples, *i.e.*, micro-videos that users preview the thumbnails yet no "click" occurs. Therefore, how to utilize these true negative samples to explicitly model users' uninterested information becomes a crucial problem.

For the past few years, several studies have been conducted on the personalized recommendation, such as collaborative filtering based models [1, 3, 12], content-based systems [5, 15, 19, 28, 29], and hybrid methods [6, 26]. Although these methods produce promising performance on recommendation, most of them suppose users' interest as static. Inspired by this, some researches consider users' interest as dynamic when designing recommendation systems and have achieved better performance [4, 7, 20, 23, 24, 27]. They, however, overlook the diverse and multi-level characteristics of users' interest. Moreover, all the aforementioned methods commonly assume that items not been clicked by users are negative and utilize them [8, 18, 22] or sample part of them as negative samples [9, 10] to represent users' uninterested items. However, these presumed negative samples may be not truly negative, and they hence may confuse the recommendation system. As we can see, the existing studies neither consider the diverse and multi-level interest nor exploit users' true uninterested items to model the recommendation system. Therefore, they cannot be directly applied to the micro-video recommendation.

To address the aforementioned problems, in this paper, we develop an end-to-end temporAL graPh-guIded recommeNdation systEm, dubbed as ALPINE, to route micro-videos. The scheme of our proposed approach is illustrated in Figure 2. Specifically, to model users' diverse and dynamic interest, we encode users' click history information into a graph where the node refers to micro-videos in the click history and the edge between two nodes stands for the temporal relationship. Based upon this graph, we design a novel Long Short-Term Memory (LSTM) network to learn users' interest representation. Afterwards, we estimate the click probability via calculating the similarity between the users' interest representation and the embedding of the given micro-video. Considering that users' interest is multi-level, we introduce a user matrix to enhance the user interest modeling by incorporating their "like" and "follow" information. And at this step, we also get a click probability with respect to users' more precise interest information. Analogously, since we know the sequence of users' disliked micro-videos, another temporal graph-based LSTM is built to characterize users' uninterested information, and the other click probability can be estimated based on true negative samples. We can thus obtain a click probability regarding users' uninterested information. Finally, the weighted sum of the above three probability scores is set as our final prediction result.

The key contributions of this work are three-fold:

- We design a novel micro-video recommendation system by jointly modeling the sequential and diverse interest of the user. In addition, considering that users' uninterested points are also dynamic and important to micro-video recommendation, we develop a temporal graph-based LSTM network to characterize users' uninterested history as well.
- To enhance users' interest representation and further improve the recommendation performance, we introduce a user matrix to record users' multi-level interest and integrate it into our recommendation system.
- We evaluate our proposed model on two public micro-video datasets to comparatively demonstrate the superiority of our model. As a side contribution, we have released the data and codes[5] of this work to facilitate other researchers.

2 RELATED WORK

Our work is closely related to video recommendation and micro-video understanding.

2.1 Video Recommendation Systems

Recommender systems are vital in video communities, such as Youtube[6], Vimeo[7], and Veoh[8]. The exiting methods can be roughly categorized as collaborative filtering based methods [1, 3, 12], content-based methods [5, 15, 19, 28, 29], and hybrid methods [6, 26]. In terms of collaborative filtering, Baluja *et al.* [1] utilized the random walk through a co-view graph to recommend YouTube videos. Chen *et al.* [3] integrated an attention mechanism into collaborative filtering with implicit feedback and evaluated its effectiveness in multimedia recommendation. However, collaborative filtering based methods cannot well solve the cold start problem. By contrast, the content-based methods recommend videos by calculating the similarity between new videos and users' historical accessed videos. For example, Mei *et al.* [15] proposed a contextual recommendation system based on multimodal fusion and relevance feedback. With respect to the hybrid models, they aim to combine the above two methods within a unified framework. For example, the recommendation model presented in [26] generates multiple ranking lists via exploring different information sources in a multi-task framework.

[5] https://anonymous1240.wixsite.com/alpine.
[6] https://www.youtube.com/.
[7] https://vimeo.com/.
[8] https://www.veoh.com/.

Since the underline assumption of the traditional video recommendation models is that users' interest is static, therefore they cannot be applied to extract users' dynamic interest.

Recently, many models have been proposed to characterize users' dynamic preferences. These methods are in three variants: Convolutional Neural Network (CNN) based methods [23, 24], Recurrent Neural Network (RNN) based methods [7, 20], and self-attention based methods [4, 27]. As a typical example in the first category, Tuan *et al.* [24] utilized 3-D CNNs to combine session clicks and content features to generate recommendations. As for RNN based methods, Quadrana *et al.* [20] proposed the RNN based approach for session-based recommendation, which relays and evolves latent hidden states of the RNNs across user sessions . In [7], the authors proposed a dynamic RNN to model users' dynamic interest for the personalized video recommendation. Due to the high time consumption and long sequence restriction, the self-attention mechanism has been applied to recommender systems and gained impressive performance. For example, Zhou *et al.* [27] proposed an attention-based user behaviour model by considering heterogeneous user behaviours in e-commerce. Although the aforementioned methods have considered users' dynamic interest and been successfully applied to video communities, they are inadequate to handle micro-video communities due to their different characteristics. In particular, micro-video communities continuously route micro-videos to users and users click their interested ones by previewing the thumbnails; whereas traditional video communities are apt to display users' interested videos via their query information. In addition, users' interest information in micro-video communities has a multi-level structure.

2.2 Micro-video Understanding

Due to the continuously booming of micro-videos on the social network, micro-video content analysis has attracted wide attention from the academic field in recent years [14, 16, 21]. Chen *et al.* [2] presented a novel low-rank multi-view embedding learning framework to perform popularity prediction for micro-videos. To better understand the micro-videos, Liu *et al.* [13] utilized LSTMs and a CNN to respectively model the spare concept and multi-modal sequential information. And Nie *et al.* [17] enhanced the acoustic modality for the venue category estimation task.Besides, there are a few researches on the micro-video recommendation. For example, Huang *et al.* [11] fused multimodal features of micro-videos to model users' personalized interest. Chen *et al.* [4] adopted forward multi-head self-attention methods with item-level and category-level attention to model user behaviours.

To the best of our knowledge, the exiting micro-video recommendation methods only extend the traditional video recommendation to the micro-video domain and ignore the distinct characteristics, especially the multi-level interest. Our proposed micro-video recommendation system differs from the aforementioned methods mainly in two aspects: 1) We regarded users' viewing history as a temporal structure and designed a graph-based LSTM to model their diverse and dynamic interest. And 2) we considered multiple types of interactions within micro-video platforms, which reflect various degrees of interest. It is worth highlighting that the users' uninterested points are also meaningful and therefore captured in our model.

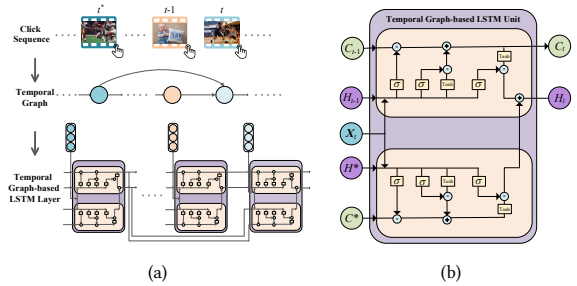


Figure 3: Structure of the temporal graph-based LSTM layer. (a) Illustration of the temporal graph construction. And (b) details of the temporal graph-based LSTM unit.

3 OUR PROPOSED MODEL

As Figure 2 illustrates, our proposed ALPINE model comprises the following three components: 1) The temporal graph-based LSTM layer. It leverages users' "click" and "not click" historical behaviours to extract their dynamic and diverse interested and uninterested feature sequences, respectively; 2) The multi-level interest modeling layer. The user matrix module explores the multi-level interest information to enhance users' interest representation; And 3) the prediction layer. It estimates users' click probability for the given micro-video.

3.1 Problem Formulation

Let v and u denote a micro-video and a user, respectively. We present the user's historical information as a sequence of micro-videos $\mathcal{U} = \{(u, v_j^t)\}_{t=1}^{m}$, where $j \in \{c, n, l, f\}$ respectively represents user's "click", "not click", "like", and "follow" behaviours, and m is the length of the sequence. As the user's interest is multi-level, its sequential behaviours can be segmented into four sub-sequences, namely "click" sequence $\mathcal{U}_c = \{(u, v_c^{t_c})\}_{t_c=1}^{m_c}$, "not click" sequence $\mathcal{U}_n = \{(u, v_n^{t_n},)\}_{t_n=1}^{m_n}$, "like" sequence $\mathcal{U}_l = \{(u, v_l^{t_l})\}_{t_l=1}^{m_l}$, and "follow" sequence $\mathcal{U}_f = \{(u, v_f^{t_f})\}_{t_f=1}^{m_f}$, where $m_c + m_n + m_l + m_f = m$. As such, the micro-video recommendation problem can be formally defined as:

Input: The user's multi-level behaviour sequences $\mathcal{U}_c$, $\mathcal{U}_n$, $\mathcal{U}_l$, and $\mathcal{U}_f$, and the given micro-video v_{new}.

Output: A recommendation system predicting the click probability of the user u on the new micro-video v_{new}.

3.2 The Temporal Graph-based LSTM Layer

To model users' dynamic interest from their historical "click" information $\mathcal{U}_c$, a direct way is to utilize the LSTM network to model the temporal sequence and obtain their interest representation. Formally, the above process can be formulated as,

$$\begin{cases} \mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o), \\ \mathbf{u}_t = tanh(\mathbf{W}_{ux}\mathbf{x}_t + \mathbf{W}_{uh}\mathbf{h}_{t-1} + \mathbf{b}_u), \\ \mathbf{c}_t = \mathbf{i}_t \odot \mathbf{u}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1}, \\ \mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t), \end{cases} \tag{1}$$

where $\mathbf{x}_t$ is the micro-video embedding at the time step t; $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{o}_t$, $\mathbf{c}_t$, and $\mathbf{h}_t$ respectively denote the input gate, forget gate, output gate, memory cell, and hidden state; σ denotes the logistic sigmoid function; and $\odot$ denotes element wise multiplication. Although the LSTM network is capable of memorizing information from sequence data, we argue that it is insufficient to capture user's diverse interest from the very long historical sequence. Particularly, if the user's historical sequence only relates to one topic, the LSTM network indeed can capture user's single interest. However, as discussed before, interest is diverse, as shown in the Figure 1. Thereby, it may fail to memorize the user's diverse interest information from the very long sequence.

To tackle the aforementioned problem, we consider to enhance the memorization of user's diverse interest by integrating an interest graph into the LSTM network. The detail of our temporal graph-based LSTM layer is illustrated in Figure 3. In particular, given the user's click sequence $\mathcal{U}_c$, we build a temporal graph $\mathcal{G}_c =< v_c, e_c >$. We view micro-videos in $\mathcal{U}_c$ as nodes, and link two nodes according to the following two rules: 1) To model the user's dynamic interest, each micro-video $v_c^{t_c}$ should be connected with its preceding micro-video $v_c^{t_c-1}$, namely $< v_c^{t_c-1}, v_c^{t_c} >$; 2) To memorize the diverse interest of the user, we force each micro-video to link with the preceding micro-videos which share the similar visual information. Given a micro-video $v_c^{t_c}$, we estimate its similarity with respect to its pre-context micro-videos and connect it with the most similar one[9], namely $< v_c^{t_c^*}, v_c^{t_c} >$. In the light of this, we construct a temporal interest graph, where each node represents one of the user's interested micro-videos, and each edge represents the relationship between the user's interested micro-videos. Moreover, for extracting the user's interested feature sequence, we design a novel graph-based LSTM network. Formally, we formulate this network as follows,

$$
\begin{cases}
\mathbf{i}_t = \sigma(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i), \\
\mathbf{f}_t = \sigma(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f), \\
\mathbf{o}_t = \sigma(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o), \\
\mathbf{u}_t = tanh(\mathbf{W}_{ux}\mathbf{x}_t + \mathbf{W}_{uh}\mathbf{h}_{t-1} + \mathbf{b}_u), \\
\mathbf{c}_t = \mathbf{i}_t \odot \mathbf{u}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1}, \\
\mathbf{i}_t^* = \sigma(\mathbf{W}_{ix}^*\mathbf{x}_t + \mathbf{W}_{ih}^*\mathbf{h}^* + \mathbf{b}_i^*), \\
\mathbf{f}_t^* = \sigma(\mathbf{W}_{fx}^*\mathbf{x}_t + \mathbf{W}_{fh}^*\mathbf{h}^* + \mathbf{b}_f^*), \\
\mathbf{o}_t^* = \sigma(\mathbf{W}_{ox}^*\mathbf{x}_t + \mathbf{W}_{oh}^*\mathbf{h}^* + \mathbf{b}_o^*), \\
\mathbf{u}_t^* = tanh(\mathbf{W}_{ux}^*\mathbf{x}_t + \mathbf{W}_{uh}^*\mathbf{h}^* + \mathbf{b}_u^*), \\
\mathbf{c}_t^* = \mathbf{i}_t^* \odot \mathbf{u}_t^* + \mathbf{f}_t^* \odot \mathbf{c}^*, \\
\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{c}_t) + \mathbf{o}_t^* \odot tanh(\mathbf{c}_t^*),
\end{cases}
\tag{2}
$$

where $\mathbf{x}_t$ is the micro-video embedding at the time step t, $\mathbf{h}_{t-1}$ and $\mathbf{c}_{t-1}$ are respectively the hidden state and memory cell at the time step $t-1$, linking by edge $< v_c^{t_c-1}, v_c^{t_c} >$, and $\mathbf{h}^*$ and $\mathbf{c}^*$ are the hidden state and memory cell at the time step t^*, linking by edge $< v_c^{t_c^*}, v_c^{t_c} >$. Therefore, our temporal graph-based LSTM network can simultaneously leverage user's neighbor and cross-time interested context information to enhance the memorization of diverse interest and further strengthen the interest representation.

[9] The number of similar micro-videos that should be linked is detailed in Section 4.6.

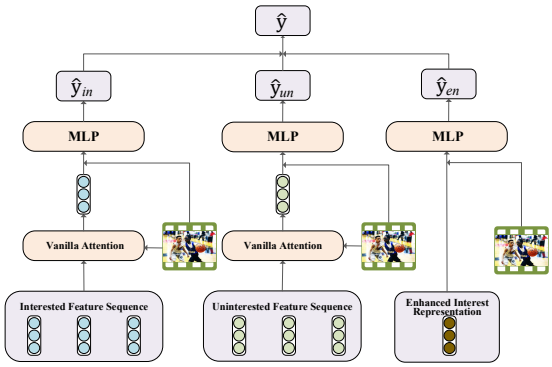


Figure 4: Structure of the Prediction Layer.

And we can obtain the user's interested feature sequence $\mathbf{F}_{in} = [\mathbf{h}_{in,1}, \mathbf{h}_{in,2}, \ldots, \mathbf{h}_{in,m_c}] \in \mathbb{R}^{d_c \times m_c}$, where d_c is the dimension of each hidden state in $\mathbf{F}_{in}$.

As the user's uninterested points are also dynamic and diverse, we build another temporal graph-based LSTM layer to model the user's $\mathcal{U}_n$ sequence and then obtain the uninterested feature sequence of the user, *i.e.*, $\mathbf{F}_{un} = [\mathbf{h}_{un,1}, \mathbf{h}_{un,2}, \ldots, \mathbf{h}_{un,m_n}] \in \mathbb{R}^{d_n \times m_n}$, where d_n is the dimension of each hidden state in $\mathbf{F}_{un}$.

3.3 The Multi-level Interest Modeling Layer

Since there are multiple interactions between a user and a micro-video and they reflect different degrees of user's interest, we propose a multi-level interest modeling layer to further obtain the enhanced interest representation. As the "like" and "follow" behaviours indicate users' stronger interest compared with the "click" one, we hence utilize the "like" and "follow" information to enhance the interest representation. Particularly, for the user u, we set the weighted sum of micro-video representations in $\mathcal{U}_l$ and $\mathcal{U}_f$ as the user's enhanced interest feature $\mathbf{f}_{en}$, formulated as,

$$
\mathbf{f}_{en} = w_l \sum_{t_l=1}^{m_l} \mathbf{x}_l^{t_l} + w_f \sum_{t_f=1}^{m_f} \mathbf{x}_f^{t_f}, \tag{3}
$$

where $\mathbf{x}_l^{t_l}$ is the embedding of micro-video $v_l^{t_l}$ in $\mathcal{U}_l$, $\mathbf{x}_f^{t_f}$ is the embedding of micro-video $v_f^{t_f}$ in $\mathcal{U}_f$, w_l and w_f are the hyper parameters controlling the weights between "like" and "follow".

With the enhanced interest representation $\mathbf{f}_{en}$, we can construct an embedding matrix $\mathbf{U} \in \mathbb{R}^{N \times D}$, *i.e.*, user matrix, where N and D respectively denote the number of users and the dimension of the enhanced interest representations. As the user's "like" and "follow" information more precisely indicates the user's interest, we can obtain more accurate interest representations using the user matrix. The user matrix $\mathbf{U}$ will be updated in the training phrase. Moreover, for each user, we utilize embedding lookup strategy to search the user's enhanced interest representation from the matrix $\mathbf{U}$ during the training and testing phrase.

3.4 The Prediction Layer

Standing on the shoulder of the user's interested feature sequence $\mathbf{F}_{in}$, uninterested feature sequence $\mathbf{F}_{un}$, and enhanced interest representation $\mathbf{f}_{en}$, we place a prediction layer to get the click probability of the given micro-video v_{new}, as shown in Figure 4. Specifically, we first feed $\mathbf{F}_{in}$ and the embedding of the given micro-video $\mathbf{x}_{new}$

into a vanilla attention layer to obtain the improved interested representation $\mathbf{f}_{in}$. Formally, the attention layer is defined as follows,

$$\begin{cases} \alpha_j = \dfrac{\exp(f(\mathbf{h}_{in,j}, \mathbf{x}_{new}))}{\sum_{j=1}^{m_c} \exp(f(\mathbf{h}_{in,j}, \mathbf{x}_{new}))}, \\ f(\mathbf{h}_{in,j}, \mathbf{x}_{new}) = \mathbf{h}_{in,j}^T \mathbf{W} \mathbf{x}_{new}, \end{cases} \tag{4}$$

where $\mathbf{h}_{in,j} \in \mathbb{R}^{d_c}$, $\mathbf{x}_{new} \in \mathbb{R}^D$, $\mathbf{W} \in \mathbb{R}^{d_c \times D}$, and α_j denotes the attention score of the j^{th} interested feature in $\mathbf{F}_{in}$. With the attention weight α_j , the improved interested representation is computed as follows,

$$\mathbf{f}_{in} = \sum_{j=1}^{m_c} \alpha_j \mathbf{h}_{in,j}. \tag{5}$$

Thereafter, we concatenate the improved interested representation $\mathbf{f}_{in}$ and the representation of the new micro-video x_{new}, and then feed it into a multi-layer perception (MLP) network, as follows,

$$\begin{cases} \mathbf{f}_1 = \phi(\mathbf{W}_1[\mathbf{f}_{in}, \mathbf{x}_{new}] + \mathbf{b}_1), \\ \hat{y}_{in} = \mathbf{W}_2 \mathbf{f}_1 + b_2, \end{cases} \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_c' \times (d_c + D)}$ and $\mathbf{W}_2 \in \mathbb{R}^{1 \times d_c'}$ denote the weight matrixes, $\mathbf{b}_1 \in \mathbb{R}^{d_c'}$ and b_2 respectively denote the bias vector and the bias value, and ϕ denotes the ReLU activation function. $\hat{y}_{in}$ is the click probability calculated by the improved interested representation $\mathbf{f}_{in}$.

Similarly, we can obtain the improved uninterested representation $\mathbf{f}_{un}$ based on $\mathbf{F}_{un}$ and $\mathbf{x}_{new}$ using another vanilla attention layer. Afterwards, we feed the concatenation of the improved uninterested representation $\mathbf{f}_{un}$ and the new micro-video embedding $\mathbf{x}_{new}$ into two MLP layers, and obtain the click probability based on the improved uninterested representation, *i.e.*, $\hat{y}_{un}$. Analogously, the click probability based on the enhanced interest representation, *i.e.*, $\hat{y}_{en}$, can be obtained by feeding the concatenation of the enhanced interest representation $\mathbf{f}_{en}$ and the new micro-video embedding $\mathbf{x}_{new}$ into two MLP layers.

Finally, the weighted sum of the above three probability values is set as our prediction result,

$$\hat{y} = \alpha_1 \hat{y}_{in} + \alpha_2 \hat{y}_{un} + \alpha_3 \hat{y}_{en}, \tag{7}$$

where α_1, α_2, and α_3 are the hyper parameters controlling the weights of $\hat{y}_{in}$, $\hat{y}_{un}$, and $\hat{y}_{en}$, respectively, and $\hat{y}$ is the final output of our model denoting the click probability of the given user on the given new micro-video.

Our method is trained as an end-to-end deep learning model equipped with the sigmoid cross-entropy loss:

$$L(\hat{y}) = -(y \log \sigma(\hat{y}) + (1 - y) \log(1 - \sigma(\hat{y}))), \tag{8}$$

where σ denotes the sigmoid activation function and $y \in \{0, 1\}$ is the ground truth that indicates whether the user clicks the micro-video or not. Besides, the Back-Propagation Through Time (BPTT) method is adopted to train our ALPINE model.

Table 1: Statistics of the two datasets.

Dataset	Dataset I	Dataset II
# users	10,000	10,986
# items	3,239,534	1,704,880
# interactions	13,661,383	12,737,619
# interaction types	4	2
#average interactions per user	1366.14	1159.44
#average interactions per item	4.28	7.47
#average clicked items per user	277	218
#interactions in training set	10,931,092	8,970,310
#interactions in test set	2,730,291	3,767,309

4 EXPERIMENTS

4.1 Dataset

Dataset I. The first dataset is released by the Kuaishou Competition[10] in China MM 2018 conference, which aims to infer users' click probabilities for new micro-videos. In this dataset, there are multiple interactions between users and micro-videos, such as "click", "not click", "like", and "follow". Particularly, "not click" means the user did not click the micro-video after previewing its thumbnail. Moreover, each behaviour is associated with a timestamp, which records when the behaviour happens. We have to mention that the timestamp has been processed such that the absolute time is unknown, but the sequential order can be obtained according to the timestamp. For each micro-video, the contest organizers have released its 2,048-d visual embedding of its thumbnail. Among the large-scale dataset, we randomly selected 10,000 users and their 3,239,534 interacted micro-videos to construct the Dataset I.

Dataset II. The second dataset is constructed by [4] for micro-video click-through prediction. It consists of 10,986 users, 1,704,880 micro-videos, and 12,737,619 interactions. Different from Dataset I, Dataset II only contains the "click" and "not click" behaviours. In this dataset, each micro-video is represented by the 512-d visual embedding extracted from its thumbnail and associated with a category label, and each user's behaviour is linked with a processed timestamp.

The statistics of the above two datasets are summarized in Table 1. The reported experimental results in this paper are based on these two datasets. Specifically, we set the first 80% of a user's historical accessed micro-videos as the training set and the rest of 20% as the test set in the Dataset I. As for Dataset II, we utilized the same setting with [4]. It is worth mentioning that we adopted the Principal Component Analysis (PCA) [25] to reduce the micro-video's visual embedding to 64 dimension.

4.2 Experimental Settings

Evaluation Protocols. To thoroughly measure our model and the baselines, we employed P@K, R@K, F@K, and Area Under Curve (AUC) as the evaluation metrics to measure the model performance from different angles. Given the recommendation list computed based on the click probability, P@K indicates the percentage of actually clicked items in the top K items of the recommendation list, R@K is the recall value of the top K items, and F@K is the harmonic average of precision and recall of the top K items. Moreover, the

[10] https://www.kesci.com/home/competition/5ad306e633a98340e004f8d1.

Table 2: Performance comparison between our proposed model and several state-of-the-art baselines over two datasets. And statistical significance over AUC between ALPINE and the best baseline (*i.e.*, THACIL) is determined by a t-test (△ denotes p-value<0.01).

Methods	Dataset I				Dataset II			
	AUC	P@50	R@50	F@50	AUC	P@50	R@50	F@50
BPR	0.595	0.290	0.387	0.331	0.583	0.241	0.181	0.206
LSTM-R	0.713	0.316	0.420	0.360	0.641	0.277	0.205	0.236
CNN-R	0.719	0.312	0.413	0.356	0.650	0.287	0.214	0.245
ATRank	0.722	0.322	0.426	0.367	0.660	0.297	0.221	0.253
NCF	0.724	0.320	0.420	0.364	0.672	0.316	0.225	0.262
THACIL	0.727	0.325	0.429	0.369	0.684	**0.324**	0.234	0.269
ALPINE	**0.739**$^{\triangle}$	**0.331**	**0.436**	**0.376**	**0.713**$^{\triangle}$	0.300	**0.460**	**0.362**

AUC is formulated as,

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{I}_u^+||\mathcal{I}_u^-|} \sum_{i \in \mathcal{I}_u^+} \sum_{j \in \mathcal{I}_u^-} \delta(\hat{y}_{u,v_i} > \hat{y}_{u,v_j}), \quad (9)$$

where $\hat{y}_{u,v_i}$ and $\hat{y}_{u,v_j}$ are the predicted probabilities of user-video pairs (u, v_i) and (u, v_j) in the test set, $\mathcal{U}$ is the user set, $\mathcal{I}_u^+$ and $\mathcal{I}_u^-$ respectively denote the set of micro-videos clicked/not clicked by the user, and δ denotes the indicator function.

Implementation Details. In the Dataset I, we utilized the 64-d visual embedding to represent the micro-video. As for the Dataset II, the concatenation of the 64-d category embedding and the 64-d visual embedding is set as the micro-video embedding. The length of users' historical sequence is set to 300. If it exceeds 300, we truncated it to 300; otherwise, we padded it to 300 and masked the padding in the network. We optimized the parameters using Adam with the initial learning rate 0.001, and the batch size is 2048. And α_1, α_2, and α_3 are 0.58, 0.18, and 0.24 on Dataset I and 0.68, 0.32, and 0 on Dataset II.

4.3 Baselines

To demonstrate the effectiveness of our proposed ALPINE model, we compared it with the following state-of-the-art methods:

- **BPR** [22]: This is a Bayesian personalized ranking model, which trains on pairwise items by maximizing the difference between the posterior probability of the positive samples and the negative ones.
- **CNN-R**: This model is a CNN based recommendation system, which utilizes the CNN structure to model sequential information. In particular, it first applies different convolutional kernels to the sequential feature matrix. Explicitly, the window size varies from one to ten, and each kernel size has 32 linear filters. Thereafter, it feeds the obtained feature map into the max pooling layer followed by a fully connected layer to obtain interest embedding. Finally, a MLP is followed to predict the click probability.
- **LSTM-R**: This model utilizes the LSTM network to model the user's sequential information. Having obtained the hidden states, it feeds them into a fully connected layer to generate the interest representation, and then a MLP module is adopted to predict the click probability.
- **ATRank** [27]: It is an attention-based user behaviour modeling framework, which captures the user's behaviour interactions in multiple semantic spaces by the self-attention mechanism.
- **NCF** [8]: It is a collaborative filtering based deep recommendation model, which learns the user embedding and the item embedding with a shallow network (element-wise product between user and item) and a deep network (concatenation of the user and item embedding followed by several MLP layers).
- **THACIL** [4]: It is a self-attention based method for the micro-video recommendation, which utilizes a multi-head self-attention layer to capture the long-term correlation within user behaviours and the item and category two level attention layer to model the fine-grained profiling of the user interest.

It is worth mentioning that THACIL and ATRank utilize the same click probability prediction layer as our model. As to the other methods including CNN-R, LSTM-R, BPR, and NCF, we fed the interest representations and the embedding of the new micro-video into the MLP layer to predict the click probability.

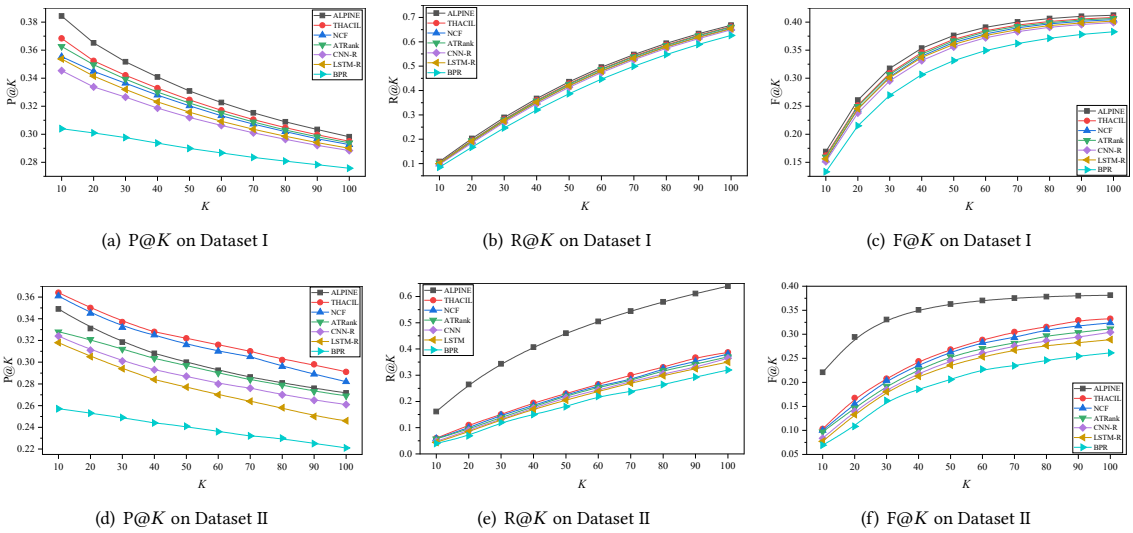
4.4 Overall Comparison

We conducted an empirical study to investigate whether our proposed model can achieve better recommendation performance. The results of all methods on two datasets are summarized in Table 2. And several observations stand out:

- BPR performs worse than the other baselines since it overlooks the sequential characteristic of the users' interest information. It hence fails to exploit the user's dynamic interest, revealing the necessity of modeling the historical sequence.
- Sequential modeling methods, including LSTM-R, CNN-R, ATRank, and THACIL, surpass the BPR model. This verifies the effectiveness of sequence modeling. Moreover, the self-attention based models, *i.e.*, ATRank and THACIL, outperform CNN-R and LSTM-R, especially the latter one. It reveals that simply utilizing the LSTM network is insufficient to capture the users' dynamic and diverse interest information from a very long sequence. The attention mechanism can implicitly reduce the memorization length by focusing on the key interest information, that is why ATRank and THACIL achieve better performance on two datasets.
- While NCF does not model the user's historical information as a sequence, it also achieves promising performance compared with the other baselines. Probably because setting a user embedding matrix and updating it in the training stage can improve the interest representation. Moreover, two operations, the element wise product and several MLPs, model the relationship between users and items better.

Table 3: Component-wise validation of our proposed ALPINE model by disabling one component each time. And statistical significance over AUC among all baselines is determined by a t-test ($\triangle$ denotes p-value<0.01 and $\diamond$ denotes p-value<0.05).

Methods	Dataset I				Dataset II			
	AUC	P@50	R@50	F@50	AUC	P@50	R@50	F@50
ALPINE_u	$0.737^{\diamond}$	0.330	0.435	0.375	$0.702^{\triangle}$	0.294	0.454	0.356
ALPINE_m	$0.735^{\triangle}$	0.329	0.433	0.374	-	-	-	-
ALPINE_um	$0.734^{\triangle}$	0.327	0.432	0.372	-	-	-	-
ALPINE_umg/ALPINE_ug	$0.716^{\triangle}$	0.318	0.426	0.363	$0.654^{\triangle}$	0.291	0.219	0.250
ALPINE	**0.739**	**0.331**	**0.436**	**0.376**	**0.713**	**0.300**	**0.460**	**0.362**



(a) P@K on Dataset I (b) R@K on Dataset I (c) F@K on Dataset I

(d) P@K on Dataset II (e) R@K on Dataset II (f) F@K on Dataset II

Figure 5: Recommendation performance versus the number of returned items K.

- ALPINE achieves the best performance, substantially surpassing all the baselines. Particularly, ALPINE presents consistent improvements over sequential models like ATRank and THACIL, verifying the importance of memorizing the prior interested information and employing the temporal graph-based LSTM network on enhancing the interest representation. In addition, our proposed ALPINE exceeds NCF, because NCF randomly initializes the user matrix rather than explores its multi-level interest information. This justifies the effectiveness of our proposed multi-level interest modeling module. Moreover, as ALPINE also characterizes the user's uninterested cues, which can further improve the recommendation performance.

In addition, we also conducted the significance test between our model and the most competitive baseline THACIL. We can see that the advantage of our model is statistically significant as p-value is 2.81×10^{-5} on the Dataset I and 4.70×10^{-6} on the Dataset II.

To justify the robustness of our proposed model, we comparatively explored the performance of our model and the baselines by varying the number of returned items K. Figure 5 shows the results regarding the performance comparison on K:

- Jointly analyzing the performance of the models in Figures 5(a) and 5(d), we found that increasing the number of returned items K degrades the precision value of the recommendation. But our model ALPINE outperforms others under the same experimental setting, especially on the Dataset I.
- The performance of all these methods over recall and F value rises fast as the number of returned items K linearly increases. Their curves then gradually ascend to a steady state. Our method ALPINE consistently and remarkably outputs a higher accuracy as compared to that of other methods, especially on the Dataset II. This verifies the robustness of our model.

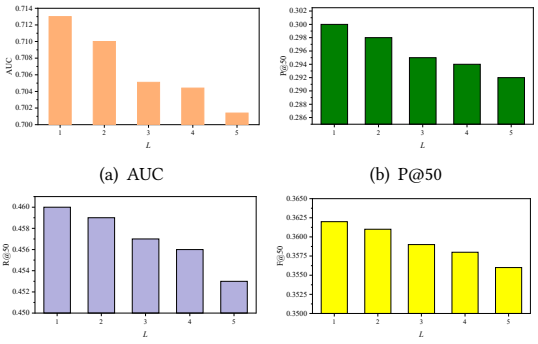
4.5 Component-wise Evaluation of ALPINE

We studied the variants of our model to further investigate the effectiveness of the uninterested representation modeling, user-matrix, and temporal interest graph:

- **ALPINE_u**: We eliminated the uninterested representation modeling part from the model. Namely, we computed the final click probability by interested representation and multi-level interest representation.
- **ALPINE_m**: We eliminated the multi-level interest module. That is, the final click probability is computed by the user's interested and uninterested representation.
- **ALPINE_um**: We only utilized the user's interested sequence to predict the click probability, namely we eliminated both the uninterested representation modeling and the multi-level interest modeling layer.
- **ALPINE_umg**: We eliminated the graph information from the ALPINE_um model.

We compared these variants on the two datasets, and Table 3 summarizes the results regarding the component-wise comparison. By jointly analyzing Table 3, we gained the following insights:

- By jointly analyzing the performance of ALPINE_u on the two datasets, it can be seen that removing the uninterested representation modeling degrades the recommendation results. To be more specific, ALPINE_u has dropped by 0.2% on the Dataset I and 1.1% on the Dataset II in terms of AUC. This verifies the effectiveness of the uninterested representation modeling.
- ALPINE surpasses ALPINE_m, indicating that incorporating the user matrix layer is beneficial to strengthen the interest representation. Moreover, compared with ALPINE_u, the performance of ALPINE_um conformably drops 0.3% under four metrics, which further reflects the effectiveness of our multi-level interest modeling layer. It is worth mentioning that the Dataset II only contains "click" and "not click" interaction, therefore the corresponding results are vacant.
- ALPINE_um shows the consistent improvements over the ALPINE_umg on the Dataset I and ALPINE_ug on the Dataset II. Specifically, the improvements of ALPINE_um over these models in terms of AUC are 2.3% on the Dataset I and 5.9% on the Dataset II, demonstrating the great advantage of our novel temporal graph-based LSTM network on capturing both dynamic and diverse interest.



(a) AUC (b) P@50

(c) R@50 (d) F@50

Figure 6: Illustration of the neighbor size L of the temporal graph on our recommendation performance.

4.6 Justification of the Temporal Graph

Apart from achieving the superior performance, the key advantage of ALPINE over other methods is that its temporal graph structure is able to strengthen the interest representation. Towards this end, we carried out experiments over the two datasets to verify the influence of the neighbor size L of the temporal graph.

In this experiment, we selected the top L similar micro-videos from the graph as neighbors of the given micro-video rather than considering the top one. Specifically, we set the average of the top L similar micro-videos' hidden states and memory cells as h^* and c^* in Eqn.(2), respectively. The comparison results versus the neighbor size L are illustrated in Figure 6. We found that the performance consistently drops under different evaluation metrics when L increases, especially the AUC drops significantly. This may be due to the fact that much more noise is introduced when a micro-video is connected with many others. Therefore, in this paper, we set L equals to one.

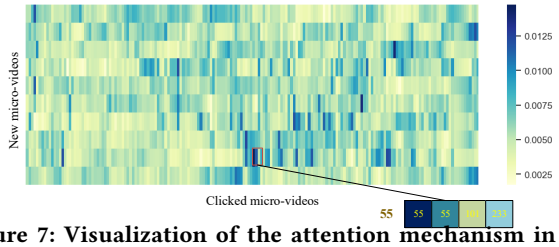


Figure 7: Visualization of the attention mechanism in the prediction layer.

4.7 Attention Visualization

As analyzed before, we fed the interested feature sequence $\mathbf{F}_{in}$ and a new micro-video's embedding $\mathbf{x}_{new}$ into a vanilla attention layer to obtain the improved interested representation. To intuitively illustrate the attention results, we randomly selected some new micro-videos from the test data and visualized the attention scores in Figure 7. Several interesting observations stand out:

- For each new micro-video, the attention scores of its historical clicked micro-videos are different, which indicates that different micro-videos in the historical sequence contribute differently.
- By and large, the earlier a micro-video locates in the sequence, the smaller the attention score is, which indicates that the latter clicked micro-videos contribute more to the recommendation. This observation strongly supports that the user's interest is dynamic.
- By visualizing the categories of micro-videos, we noticed that, micro-videos from the same category contributes more to the recommendation results. As shown in the sub-figure of Figure 7, the new given micro-video belongs to the 55-th category, and the attention mainly focuses on the micro-videos of the same category in the historical sequence. This demonstrates the attention layer can help obtain improved features according to different new micro-videos.

5 CONCLUSION

In this work, we present a temporal graph-based LSTM model to intelligently route micro-videos to the target users. To capture the users' dynamic and diverse interest, we encode their historical interaction sequence into a temporal graph and then design a novel temporal graph-based LSTM to model it. As different interactions reflect different degrees of interest, we build a multi-level interest modeling layer to enhance users' interest representation. Moreover, our model extracts uninterested information from true negative samples to improve the recommendation performance. To justify our scheme, we perform extensive experiments on two public datasets, and the experimental results demonstrate the effectiveness of our model.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.:61702300, No.:61702302, No.: 61802231, and No. U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.: ZR2019JQ23, No.:ZR2019QF001; the Future Talents Research Funds of Shandong University, No.: 2018WLJH 63.

REFERENCES

[1] Shumeet Baluja, Rohan Seth, D Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. 2008. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the ACM International Conference on World Wide Web*. 895–904.

[2] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: predicting the popularity of micro-videos via a transductive model. In *Proceedings of the ACM International Conference on Multimedia*. 898–907.

[3] 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 International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–344.

[4] 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 ACM International Conference on Multimedia*. 1146–1153.

[5] 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 ACM International Conference on Multimedia*. 597–606.

[6] Andrea Ferracani, Daniele Pezzatini, Marco Bertini, and Alberto Del Bimbo. 2016. Item-based video recommendation: An hybrid approach considering human factors. In *Proceedings of the ACM on International Conference on Multimedia Retrieval*. 351–354.

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

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

[9] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 549–558.

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

[11] Lei Huang and Bin Luo. 2017. Personalized micro-Video recommendation via hierarchical user interest modeling. In *Springer Pacific Rim Conference on Multimedia*. 564–574.

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

[13] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the ACM International Conference on Multimedia*. 970–978.

[14] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2018. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning. *IEEE Transactions on Image Processing* 28, 3 (2018), 1235–1247.

[15] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems* 29, 2 (2011), 10.

[16] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. Learning from multiple social networks. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 2 (2016), 1–118.

[17] 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 ACM International Conference on Multimedia*. 1192–1200.

[18] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N Liu, Rajan Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *Proceedings of the IEEE International Conference on Data Mining*. 502–511.

[19] Jonghun Park, Sang-Jin Lee, Sung-Jun Lee, Kwanho Kim, Beom-Suk Chung, and Yong-Ki Lee. 2011. Online video recommendation through tag-cloud aggregation. *IEEE Transaction on MultiMedia* 18, 1 (2011), 78–87.

[20] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 130–137.

[21] Miriam Redi, Neil O'Hare, Rossano Schifanella, Michele Trevisiol, and Alejandro Jaimes. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4272–4279.

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

[23] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 565–573.

[24] Trinh Xuan Tuan and Tu Minh Phuong. 2017. 3D convolutional networks for session-based recommendation with content features. In *Proceedings of ACM International Conference on Recommender Systems*. 138–146.

[25] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1-3 (1987), 37–52.

[26] 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 ACM International Conference on Multimedia*. 1521–1524.

[27] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*.

[28] Xiangmin Zhou, Lei Chen, Yanchun Zhang, Longbing Cao, Guangyan Huang, and Chen Wang. 2015. Online video recommendation in sharing community. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1645–1656.

[29] Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. 2013. Videotopic: Content-based video recommendation using a topic model. In *Proceedings of the IEEE International Symposium on Multimedia*. 219–222.