2019 IEEE International Conference on Multimedia and Expo (ICME)

HERDING EFFECT BASED ATTENTION FOR PERSONALIZED TIME-SYNC VIDEO RECOMMENDATION

Wenmian Yang[1,2], *Wenyuan Gao*[1], *Xiaojie Zhou*[1], ✉*Weijia Jia* [2,1], *Shaohua Zhang*[1,2], *Yutao Luo*[1]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]State Key Lab of IoT for Smart City, CIS, University of Macau, Macau, SAR China
sdq11111@sjtu.edu.cn, jiawj@um.edu.mo

ABSTRACT

Time-sync comment (TSC) is a new form of user-interaction review associated with real-time video contents, which contains a user's preferences for videos and therefore well suited as the data source for video recommendations. However, existing review-based recommendation methods ignore the context-dependent (generated by user-interaction), real-time, and time-sensitive properties of TSC data. To bridge the above gaps, in this paper, we use video images and users' TSCs to design an Image-Text Fusion model with a novel Herding Effect Attention mechanism (called ITF-HEA), which can predict users' favorite videos with model-based collaborative filtering. Specifically, in the HEA mechanism, we weight the context information based on the semantic similarities and time intervals between each TSC and its context, thereby considering influences of the herding effect in the model. Experiments show that ITF-HEA is on average 3.78% higher than the state-of-the-art method upon F1-score in baselines.

***Index Terms*—** Recommendation System, Collaborative Filtering, Herding Effect, Data Mining

1. INTRODUCTION

Recently, watching online videos of news and amusement has become mainstream entertainment during people's leisure time. Therefore, efficient and accurate personalized video recommendation methods bring significant convenience to their life. Most of the video recommendation methods focus on users' behaviors such as their browsing history [1, 2] and reviews [3, 4]. In real scenarios, however, most people are unwilling to do high-quality reviews after watching videos, which causes the scarcity of valuable video reviews. Furthermore, some multi-feature based methods [5, 6] combine image information with review information to generate users' interests from a more comprehensive perspective. However, their methods have only achieved limited improvement because the images and reviews usually contain unequal information [7]. That is, the text information and image information generally describe the different amount of contents. An image in a video only describes one moment of the video content, while a review usually describes the overall contents of the video. The information gap causes the fusion of reviews and images to lose great information.

Meanwhile, a new form of user-interactive review – time-sync comment (TSC) (first introduced by Wu *et al.* [8], see Fig. 1) has become increasingly popular in China and Japan, especially among young people. Nowadays, many popular Chinese video websites such as Youku (http://www.youku.com), Bilibili (http://bilibili.tv) and the Japanese video website NICONICO (http://www.nicovideo.jp) support the TSC. TSCs convey information involving the content of the current video frame, feelings of users or replies to other TSCs, which can accurately express the users' preferences for the video. Moreover, each TSC has a corresponding timestamp to record the posted time. Compared with traditional video reviews, TSCs are much easier to obtain their corresponding images by timestamps. The users' real-time feedbacks and the vast amount make TSCs valuable and accessible sources for personalized video recommendations.

In this paper, we focus on mining the users' preferences and videos' features from TSCs and corresponding images to recommend videos towards users through model-based Collaborative Filtering (CF). TSCs have several features distinguished from the traditional video reviews: ***(1)Context-dependent.*** TSCs are usually context-dependent, i.e., the latter comments often depend on the former ones. This phenomenon is known as the herding effect in social science [9, 10]. An example of the herding effect is shown in Fig. 1. User A said "I love the male commander!" to express his love to the role male commander when he appears in the video. After a few seconds, user B and user C followed up by saying "I like the male commander too..." and "I am so sad when he died." In this case, user B and user C may not make their comments if user A has not. That is, the emergence of a TSC is usually not independent, but a probability event influenced by other preorder comments. ***(2)Real-time.*** Each TSC has a timestamp synchronous to the playback time of the video. The coverage of a TSC is usually only a short time before its timestamp. Therefore, the content of each TSC is closely related to the video content corresponding to its timestamp, which makes it easy to sample corresponding image information by timestamp. ***(3)Time-sensitive.*** According to our observation, TSCs with a large interval of timestamps are unlikely to discuss the same topic, even if their semantics are similar. Users are more likely to follow those newer TSCs than older ones. As a result, the herding effect mentioned before will not last long. These features make TSC a particular review. However, most of the current TSC-based recommendations [3, 11] assume that TSCs are independent of each other and ignore the time information. Such assumption ignores the above features of TSCs which causes the loss of crucial semantic information and affects the accuracy of results. Therefore, how to take TSCs' features of context-dependent (herding effect), real-time and time-sensitive into account to extract the textual information and fuse it with visual information accurately and effectively are the central challenges.

Based on the above motivations and challenges, we propose an Image-Text Fusion model with a novel Herding Effect Attention mechanism (called ITF-HEA). In ITF-HEA, We generate users' preferences and summarize video contents through model-based CF. To analyze the influence of text information, image information and contextual information separately, we split ITF-HEA into two models: Text-based Model(TM) and Image-Text Fusion model (ITF), and one attention mechanism: Herding Effect Attention (HEA) mechanism. Specifically, in TM, we sample and embed the TSCs to obtain the sentence vectors (TSC features) by bidirectional Long Short-Term Memory (LSTM) at first, and then combine TSC features with the hidden (embedding) features of the users and videos respectively to predict the likeness of the user to the video. In ITF, we sample the corresponding video frame (image) features and incorporate them with TSC features to replace single TSC features in

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00085

TM. Finally, we design the HEA mechanism which is based on contextual semantic similarity and time interval of TSCs to incorporate contextual information into TSC features to replace the features in TM and ITF.

The main contributions of this paper are as follows:

1 We propose a novel HEA mechanism, which takes TSCs' features of contextual relevance, real time and timeliness into account, to extract the textual features of the TSCs more accurately and effectively.

2 We design an Image-Text Fusion model using model-based CF, combine it with HEA mechanism and get the ITF-HEA, which can predict the likeness of the user to the video more accurately and sufficiently.

3 We evaluate the ITF-HEA with real-world datasets on mainstream video-sharing websites and compare it with state-of-the-art video recommendation methods. The results show that our model outperforms baselines in both precision and F1-score on average 3.3% and 3.78% respectively.

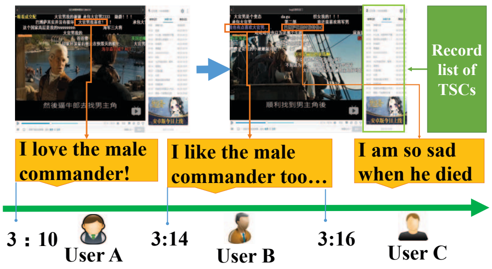


Fig. 1. Examples of TSCs.

2. RELATED WORK

In this section, we discuss the related work in three aspects.

Time-sync video comments are first introduced by Wu *et al.* [8]. Then, Yang *et al.* [12] sum up the features of TSCs, which inspire our work. LV *et al.* [13] propose a video understanding framework to assign temporal labels to highlighted video shots. They are the first to analyze the TSCs using the neural network. Recently, Liao *et al.* [14] present a larger-scale TSC dataset with four-level structures and rich self-labeled attributes, which brings convenience for future research on TSCs. The above methods show that TSC is a kind of data with great potential and development value.

Video recommendation has attracted great attention from both the industry and academia. Most of the current state-of-the-art methods are based on CF. Mcauley *et al.* [15] combine latent rating dimensions with latent review topics, which is a review-based method. Diao *et al.* [16] propose a probabilistic model based on CF and topic modeling, which is an LDA [17] based method and allows capturing the interest distribution of users and the content distribution of movies. He *et al.* [18] propose a scalable factorization model to incorporate visual signals into predictors of people's opinions, which is the state-of-the-art visual-based model. However, the above recommendation methods are not well-designed for TSCs as they ignore the interactive, real-time, and timeliness properties of TSC data.

Attention Mechanism has been shown effective in natural language processing [19, 20, 21]. Recently, attention models have been used increasingly in recommendation systems to assign weights to user-item pairs. Chen *et al.* [22] introduce a novel attention mechanism in CF to address the challenging item and component-level implicit feedback in the multimedia recommendation, which can be seamlessly incorporated into classic CF models with implicit feedback. Seo *et al.* [23] propose to model user preferences and item properties using convolutional neural networks (CNN) with dual local and global attention. Our herding effect attention mechanism adopts the soft attention [24], which learns the attentive weights based on the importance to the final task.

3. MODEL-BASED COLLABORATIVE FILTERING

In this section, we describe two CF models and an attention mechanism. First, the problem formulation is provided in Section 3.1. Then, we propose a Text-based Model by using textual features of TSCs in Section 3.2. Next, we design an Image-Text Fusion Model to jointly model video images as well as TSCs in Section 3.3. Finally, to take full consideration of the features of TSCs, we implement Herding Effect Attention mechanism and give the complete neural network structure of the Image-Text Fusion Model with Herding Effect Attention in Section 3.4.

3.1. Problem Formulation

Suppose there are N TSCs, $\boldsymbol{TSC} = \{tsc_1, tsc_2, ..., tsc_N\}$. For tsc_i, we define the corresponding visual feature as $\boldsymbol{vsl_i}$ (see Section 3.3 for details) and sentiment polarity pol_i which is determined by the Stanford sentiment analysis toolkit (http://nlp.stanford.edu/sentiment). Besides, we define u_i to represent the user ID and v_i to express the video ID of tsc_i.

As mentioned in section 1, TSCs are easily affected by previous comments. Therefore, for tsc_i, we continuously sample M preorder TSCs $\boldsymbol{Context_i} = \{pre_{i,1}, pre_{i,2}, ..., pre_{i,M}\}$ as context information ($pre_{i,M}$ is tsc_i itself). For each $pre_{i,j} \in \boldsymbol{Context_i}$, we define the time-stamp $t_{i,j}$ to represent its posted video time. The word list of tsc_i and its context information $pre_{i,j}$ are defined as $\boldsymbol{w_i} = \{w_i^1, w_i^2, ..., w_i^{L_i}\}$ and $\boldsymbol{w_{i,j}} = \{w_{i,j}^1, w_{i,j}^2, ..., w_{i,j}^{L_{i,j}}\}$ where L_i and $L_{i,j}$ are the length of tsc_i and $pre_{i,j}$.

Given total TSCs $\boldsymbol{WT} = \{\boldsymbol{w_i} | 1 \leq i \leq N\}$ and their corresponding sentiments $\boldsymbol{POL} = \{pol_i | 1 \leq i \leq n\}$, user ID list $\boldsymbol{U} = \{u_1, u_2, ..., u_N\}$, video ID list $\boldsymbol{V} = \{v_1, v_2, ..., v_N\}$, visual feature list $\boldsymbol{VSL} = \{vsl_1, vsl_2, ..., vsl_N\}$, and the context information $\boldsymbol{WC} = \{\{\boldsymbol{w_{1,1}, ..., w_{1,M}}\}, ..., \{\boldsymbol{w_{N,1}, ..., w_{N,M}}\}\}$ with the corresponding time-stamp $\boldsymbol{T} = \{\{t_{1,1}, ..., t_{1,M}\}, ..., \{t_{N,1}, ..., t_{N,M}\}\}$, our task is to predict the likeness of user $u \in \boldsymbol{UID}$ on video $v \in \boldsymbol{VID}$, where $\boldsymbol{UID}$ and $\boldsymbol{VID}$ are the sets that contain the unique IDs in $\boldsymbol{U}$ and $\boldsymbol{V}$. The top X videos among the final results are recommended to the corresponding user.

3.2. Text-based Model

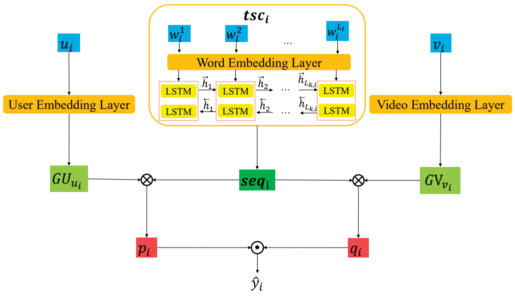


Fig. 2. Text-based Model

Intuitively, the preference of users is extracted from their published TSCs in corresponding videos, while the textual features of the videos are summed up from the TSCs published in the video. Based on above, in this section, we first extract the features of TSCs by bidirectional LSTM. Then, features of TSCs are merged with the latent factors of users and videos. Finally the likenesses of the users to videos are predicted by CF. The general framework of Text-based Model (TM) is shown in Fig. 2.

More concretely, to capture the word sequential information from TSCs, we use the Bidirectional Long Short-Term Memory (Bi-LSTM) network [25] to convert word features into TSC features.

For each tsc_i, we have

$$\overrightarrow{\boldsymbol{h}}_{\boldsymbol{t}} = LSTM(w_i^t, \overrightarrow{\boldsymbol{h}}_{\boldsymbol{t-1}}) \tag{1}$$

$$\overleftarrow{\boldsymbol{h}}_{\boldsymbol{t}} = LSTM(w_i^t, \overleftarrow{\boldsymbol{h}}_{\boldsymbol{t+1}}) \tag{2}$$

and

$$\boldsymbol{seq_i} = \frac{1}{L_i} \sum_{t=1}^{L_i} (\overrightarrow{\boldsymbol{h}}_t \oplus \overleftarrow{\boldsymbol{h}}_t) \tag{3}$$

where $\boldsymbol{seq_i} \in \mathbb{R}^d$ and $\oplus$ denotes vector concatenation. After LSTM layer, we get the sequence feature $\boldsymbol{seq_i}$ as the output.

Then, we define $\boldsymbol{GU_{u_i}}$ as the latent factor of user u_i, which is the feature based on user's historical preference. Likewise, the feature of video v_i is defined as $\boldsymbol{GV_{v_i}}$. Afterward, we design $\otimes$ function to merge $\boldsymbol{GU_{u_i}}$ and $\boldsymbol{GV_{v_i}}$ with $\boldsymbol{seq_i}$ respectively, and obtain

$$\boldsymbol{p_i} = \boldsymbol{G_{u_i}} \otimes \boldsymbol{seq_i} \tag{4}$$

and

$$\boldsymbol{q_i} = \boldsymbol{G_{v_i}} \otimes \boldsymbol{seq_i} \tag{5}$$

where $\otimes$:$R^d \times R^d - > R^d$ is an element-wise product function to merge two d dimensional vectors into one. Specifically,

$$(a_1, ..., a_d) \otimes (b_1, ..., b_d) = (a_1 b_1, ..., a_d b_d)$$

In our framework, we take the prediction of a user's favor to a video (or video clip) as a binary classification problem, where 1 means a user likes the video, and 0 otherwise. Therefore, we define the likeness of user u_i to video v_i though tsc_i in the training data as

$$\hat{y}_i = sigmoid(\boldsymbol{p_i} \odot \boldsymbol{q_i}) \tag{6}$$

where $sigmoid(x) = \frac{1}{1+e^{-x}}$ and "$\odot$" denotes inner product.

Generally, users comment on their favorite videos with positive sentiment. Therefore, we determine the polarity of each TSC by the Stanford sentiment analysis toolkit [26]. For simplicity, we set the polarity of each TSC as 1 if the result is positive or neutral, and 0 otherwise. We define $y_i = pol_i$ as the ground truth of the likeness of user u_i for video v_i through tsc_i, where pol_i is the polarity of tsc_i.

At last, we use the binary cross-entropy as our loss function to model user preference. The final objective function is maximized as:

$$L = \sum_{i=1}^{N} (y_i \cdot ln\hat{y}_i + (1 - y_i) \cdot ln(1 - \hat{y}_i)) \tag{7}$$

In the training phase, the parameters can be learned via Adam [27].

After trainning, we use

$$\hat{y}_{u,v} = \boldsymbol{GU_u} \odot \boldsymbol{GV_v} \tag{8}$$

to express the predicted likeness of user u on video v, and

$$Po_{u,v} = \frac{\sum_{i \in List_{u,v}} pol_i}{|List_{u,v}|} \tag{9}$$

to express the real likeness of user u for video v, where $List_{u,v}$ expresses the total TSCs that user u has commented on video v. Then, the ground truth of testing data is defined as

$$y_{u,v} = \begin{cases} 0 & Po_{u,v} < 0.5 \\ 1 & Po_{u,v} \geq 0.5 \end{cases} \tag{10}$$

3.3. Image-Text Fusion Model

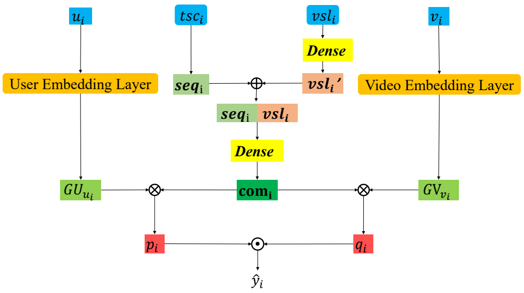


Fig. 3. Image-Text Fusion Model

In the time-sync video, each TSC has a timestamp that records the corresponding video time when the TSC is published. So that, we can easily obtain the corresponding image information for better feature extraction. In this section, we focus on merging TSC text features with corresponding visual features to obtain more comprehensive features. The general framework of Image-Text Fusion model (ITF) is shown in Fig. 3.

For tsc_i, we use vsl_i to indicate the visual feature when tsc_i is posted. The visual features are with the output of 4096-way obtained from a public TSC data set extracted by *Chen et al.* [3], which are trained by the Caffe reference model with 5 convolutional layers followed by 3 fully-connected layers that have been pre-trained on 1.2 million ImageNet (ILSVRC2010) images.

Since the dimension of $\boldsymbol{vsl_i}$ is 4096, we reduce its dimension to d and get $\boldsymbol{vsl_i'}$ by

$$\boldsymbol{vsl_i'} = Dense(\boldsymbol{vsl_i}) \tag{11}$$

where $Dense$ is the fully-connected layer with the activation function elu [28].

To combine the image features and textual features, we first concatenate the sequence feature seq_i with the visual feature vsl_i', and obtain the $2 \times d$-dimensional vector com_i:

$$\boldsymbol{com_i} = \boldsymbol{seq_i} \oplus \boldsymbol{vsl_i'} \tag{12}$$

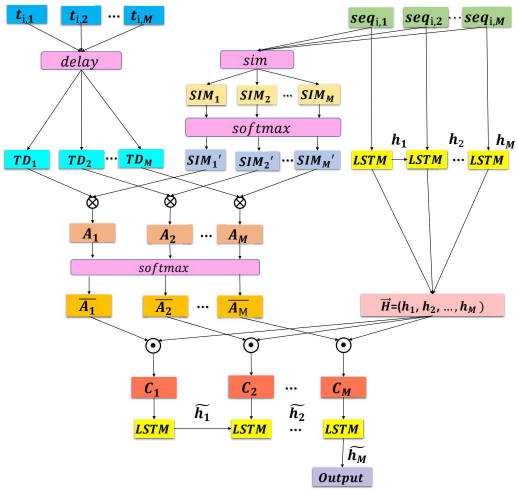
Then, we reduce the demension of com_i to d and get the com_i' by

$$\boldsymbol{com_i'} = Dense(\boldsymbol{com_i}) \tag{13}$$

Finally, we use com_i' instead of seq_i to merge with $\boldsymbol{GU_{u_i}}$ and $\boldsymbol{GV_{v_i}}$ by $Eq.(4)$ and $Eq.(5)$ and predict the likeness by $Eq.(6)$.

3.4. Herding Effect Attention

Existing review-based recommendation methods usually handle each comment separately without considering the context associations between the comments. However, TSCs are highly semantic relevant and time-related, which is so-called the herding effect. That is, TSCs may be affected by other preorder TSCs on the similar topic. Also, TSCs with similar semantics and the short interval of the time-stamp are more likely to influence each other. Based



$t_{i,1}$ $t_{i,2}$ … $t_{i,M}$
delay
$seq_{i,1}$ $seq_{i,2}$ … $seq_{i,M}$
sim
SIM_1 SIM_2 … SIM_M
softmax
TD_1 TD_2 … TD_M
SIM_1' SIM_2' … SIM_M'
h_1 h_2 h_M
LSTM LSTM … LSTM
A_1 A_2 … A_M
softmax
$\overline{A_1}$ $\overline{A_2}$ … $\overline{A_M}$
$\overline{H}=(h_1, h_2, \ldots, h_M)$
C_1 C_2 … C_M
$\widetilde{h_1}$ $\widetilde{h_2}$ $\widetilde{h_M}$
LSTM LSTM … LSTM
Output

Fig. 4. Herding Effect Attention

on above, we design an HEA mechanism, which calculates the influence weights of TSC contexts by their semantic similarities and timestamp intervals in an LSTM-based encoder-decoder framework. The framework of HEA is shown in Fig. 4.

Formally, for tsc_i, we sample M continuous preorder TSCs $\boldsymbol{Context_i} = \{pre_{i,1}, pre_{i,2}, ..., pre_{i,M}\}$ as the context information and get context features $\boldsymbol{SEQ_i} = [\boldsymbol{seq_{i,1}, ..., seq_{i,M}}]$ by $Eq.(1)$ - (3), where $t_{i,1} < t_{i,2} < .... < t_{i,M}$.

We formalize the HEA into an encoder-decoder framework. Given context features $\boldsymbol{SEQ_i}$ as the input to the LSTM, the output are obtained as:

$$\boldsymbol{h_t} = LSTM(\boldsymbol{h_{t-1}}, \boldsymbol{seq_{i,t}}) \tag{14}$$

We use $\boldsymbol{H} = (\boldsymbol{h_1, h_2, ..., h_M})$ to express the hidden status vectors of the encoder output.

To calculate the influence weights of contexts of TSCs, for $pre_{i,j}$, we define the semantic similarity vector and time delay vector as $\boldsymbol{SIM_j} = (sim(j,1), sim(j,2), ..., sim(j,M))$ and $\boldsymbol{TD_j} = (delay(j,1), delay(j,2), ..., delay(j,M))$, where

$$sim(k,j) = \frac{\boldsymbol{seq_{i,k}} \odot \boldsymbol{seq_{i,j}}}{|\boldsymbol{seq_{i,k}}||\boldsymbol{seq_{i,j}}|} \tag{15}$$

represents the semantic similarity between $pre_{i,j}$ and $pre_{i,k}$, and

$$delay(j,k) = \begin{cases} \mathrm{e}^{-\beta(t_{i,j}-t_{i,k})} & j > k \\ 0 & j <= k \end{cases} \tag{16}$$

represents the influence of $pre_{i,k}$ on $pre_{i,j}$ decreases with the increasing time interval (β is a hyper-parameter that will be discussed in Section 4).

Since the semantic similarity may have negative numbers, we first normalize $\boldsymbol{SIM_j}$ by softmax as:

$$\boldsymbol{SIM_j}' = (\frac{\mathrm{e}^{sim(j,1)}}{\sum_{k=1}^{M} \mathrm{e}^{sim_{j,k}}}, \frac{\mathrm{e}^{sim(j,2)}}{\sum_{k=1}^{M} \mathrm{e}^{sim_{j,k}}}, ..., \frac{\mathrm{e}^{sim(j,M)}}{\sum_{k=1}^{M} \mathrm{e}^{sim_{j,k}}})$$

Next, we calculate the attention score vector of $pre_{i,j}$ as:

$$\boldsymbol{A_j} = \boldsymbol{SIM_j}' \otimes \boldsymbol{TD_j} \tag{17}$$

The final attention score distribution $\overline{\boldsymbol{A_j}}$ is obtained by normalizing the attention score vector $\boldsymbol{A_j}$ by softmax function.

We compute the input of decoder as:

$$\boldsymbol{C_j} = \overline{\boldsymbol{A_j}} \odot \boldsymbol{H} \tag{18}$$

and get the output as:

$$\widetilde{\boldsymbol{h_1}} = LSTM(\boldsymbol{C_1}) \tag{19}$$

$$\widetilde{\boldsymbol{h_t}} = LSTM(\boldsymbol{C_t}, \widetilde{\boldsymbol{h_{t-1}}}) \tag{20}$$

Finally, we use $\widetilde{\boldsymbol{h_M}}$ instead of $\boldsymbol{seq_i}$ that we used in Section 3.2 and 3.3 as the textual feature.

We integrate the context-dependent, real-time and time-sensitive properties of the TSCs into the model by the HEA mechanism, which can be applied in both TM and ITF to improve the TSC feature extraction. The complete network structure of the ITF-HEA is shown in Fig. 5.

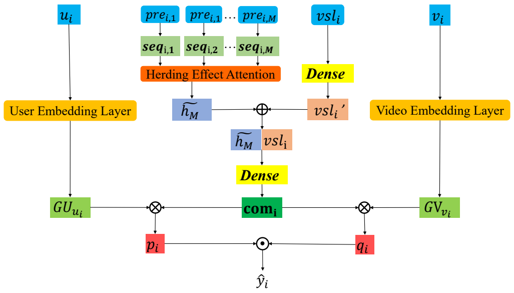


Fig. 5. Complete network structure

4. EXPERIMENTS

In this section, we demonstrate the effectiveness of our proposed method by comparing with 4 well-known methods of video recommendation. We provide necessary parameters of our model at first and then analyze the performance of our model on time-sync video recommendation. Finally, we analyze the effect of the hyper-parameters on the experimental results.

4.1. Experimental Setup and Dataset

The data used in this paper are crawled from a Chinese time-sync video site Bilibili by Chen *et al.* [3], which are obtained from the movie category till December 10th, 2015. In this paper, we select 100 users who have posted the most TSCs and commented on more than 40 videos. These users have commented on a total of 871 videos, and we select all the comments in those videos as a sub-dataset. In the sub-dataset, 423,384 users have published 1,319,475 TSCs in total.

For each of the 100 users, we select half of the videos where they have commented as the training set, and the other half as the test set. We make sure that at least 20 videos per user can be recommended (To ensure the effectiveness of top20). In the test set, we get 2,995 $(user, video)$ pairs, where 1,972 pairs are positive, and 1,023 pairs are negative in sentiment polarity.

In the training set, we obtain 2,811 $(user, video)$ pairs with 11,775 TSCs (a user may make more than one TSC in a video), where 8,124 TSCs are positive and 3,651 TSCs are negative.

In our model, hyper-parameter β and number of contextual TSCs M need to be decided. We select 35% data of the test set (actually 1,075 $(user, video)$ pairs) as the validation set to tune β. The initial learning rate of Adam [27] is 0.001 and the vector dimension d is set as 128. We get the best results when $\beta = 0.2$, and $M = 10$, which are discussed in Section 4.2.

4.2. Results

In this section, we use the test set described in Section 4.1 to compare our complete model with existing methods.

To evaluate the performance of the proposed models, we compare our model with the following methods as baselines:

- **HFT:** A state-of-the-art method regarding making rating prediction with textual reviews [15]. In the experiments, we set the ratings of positively commented videos as 1, and 0 otherwise.
- **JMARS:** A Latent Dirichlet Allocation (LDA) based method to make rating prediction with textual reviews [16].
- **VBPR:** A visual-based recommendation method [18].
- **KFRCI:** A novel Key Frame Recommender by modeling user TSCs and keyframe Images simultaneously [3]. In the experiments, the likeness score of the video is the average score of all the frames the users have commented on.
- **ITF-HEA:** The Image-Text Fusion Model proposed in Section 3.3 with Herding Effect Attention mechanism proposed in Section 3.4.

Table 1. Precision and F1-score of each method

	Top5		Top10		Top20	
	Prec	F1	Prec	F1	Prec	F1
HFK	0.856	0.3463	0.812	0.5310	0.732	0.7371
JMARS	0.878	0.3552	0.810	0.5560	0.732	0.7367
VBPR	0.892	0.3608	0.83	0.5760	0.779	0.7840
KFRCI	0.954	0.3859	0.897	0.6036	0.818	0.8233
ITF-HEA	**0.976**	**0.3948**	**0.932**	**0.6272**	**0.860**	**0.8656**

For each method in the baselines, we select a set of the best experimental parameters according to the range of the parameters given in their experiments and calculate the likenesses/ratings between users and videos.

Our experiments are conducted by predicting Top 5, 10, and 20 favorite videos respectively. The Top X is the top prediction of user's likeness to the videos in test set calculated by Eq. (8). We recommend all X videos to each user and consider these are the user's favorite videos. We adopt F1-score and precision to evaluate the performance of the baselines and our models. All the models are repeated for 10 times, and we report the average values as the final results for clear illustration.

The results of F1-score and precision are shown in Table 1. From Table 1, we can see: ITF-HEA achieves the best performance on F1 and precision (F1 is proportional to precision in the Top 5, 10 and 20). It has enhanced the performance by about 2.30%, 3.91% and 5.14% (on average 3.78%) upon F1-score and 2.20%, 3.50% and 4.20% (on average 3.30%) upon Precision on Top5, 10 and 20 respectively compared with KFRCI, which performs best among baselines. In other methods of baselines, the vision-based method VBPR has better performance than the others; the text-based method HFT and JMARS have similar performance, while the PMF method has the worst.

Next, we compare the models proposed in Section 3:

- **TM:** The Text-based Model proposed in Section 3.2.
- **T-HEA:** The Text-based Model proposed in Section 3.2 with Herding Effect Attention mechanism proposed in Section 3.4.
- **ITF:** The Image-Text Fusion Model proposed in Section 3.3.
- **ITF-HEA:** The Image-Text Fusion Model proposed in Section 3.3 with Herding Effect Attention mechanism proposed in Section 3.4.

to analyze the effects of text features, image features and the attention mechanism in our model. The results of F1-score and precision are shown in Table 2.

Table 2. Precision and F1-score of the models proposed in Section 3

	Top5		Top10		Top20	
	Prec	F1	Prec	F1	Prec	F1
TM	0.932	0.2363	0.887	0.5969	0.790	0.7950
T-HEA	0.956	0.3867	0.914	0.6151	0.817	0.8223
ITF	0.952	0.3851	0.892	0.6003	0.805	0.8107
ITF-HEA	**0.976**	**0.3948**	**0.932**	**0.6272**	**0.860**	**0.8656**

The results show that although T-HEA only uses the textual information, the experimental results are still better than ITF. T-HEA even has better performance than the state-of-the-art method KFRCI, which shows that our HEA mechanism can effectively offset the effects of the herding effect and improve the performance of the model. The results also show that the context and timestamp of the TSCs are vital information and need to be considered.

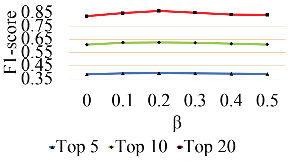


Fig. 6. The influence of the hyper-parameter β on Top 5, Top 10 and Top 20

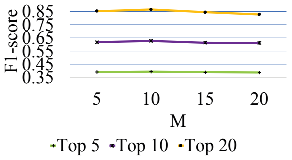


Fig. 7. The influence of the number of context M on Top 5, Top 10 and Top 20

Finally, we discuss the influence of hyper-parameter β and the number of contextual TSCs M on the experimental results. We fix $M = 10$, changing the value of β from 0 to 0.5 (the step size is 0.1), and calculate the F1-score of Top 5,10 and 20 users' favorite videos in the validation set at first. The best results are obtained when $\beta = 0.2$. We also calculate the F1-score for the different hyper-parameters in the test set, and the results are shown in Fig. 6. The hyper-parameter β gains the best performance when $\beta = 0.2$ in any case, which is the same with the validation set. When β is bigger, the result of the experiment is worse because it weakens the weight of other TSCs in the attention layer. When $\beta = 0$, it has the worst performance, because the time information is not considered.

For the number of context length M, we fix $\beta = 0.2$, and set M as 5, 10, 15 and 20, respectively. The results are shown in Fig. 7. IFT-HEA gets the best performance when $M = 10$ and the worst when $M = 20$. This result confirms that the herding effect of TSCs is time-sensitive and will not last long, which meets our observation in Section 1.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel personalized online video recommendation with the dataset of both TSCs and its corresponding images through model-based CF method. To extract the textual features of the TSCs more accurately and effectively, we designed the

HEA mechanism to add influence weight to each TSC based on their semantic similarity and time interval. In this way, we integrated the context-dependent, real-time and time-sensitive properties of TSCs in the neural network framework and predicted the users' preferences for online videos accurately and effectively. Extensive experiments on real-world dataset proved that our model could recommend videos to users more precisely than the state-of-the-art method with the HEA mechanism.

This is the first step towards our goal in personalized video recommendation, and there is much space for further improvements. For example, to design a more accurate fusion model to capture comprehensive user preference is a challenging problem. Besides, how to measure the weight of each video to the user's preference is also a challenging problem.

6. ACKNOWLEDGEMENTS

This work is supported by National China 973 Project No. 2015CB352401; Chinese National Research Fund (NSFC) Key Project No. 61532013 and No. 61872239. 0007/2018/A1, DCT-MoST Joint-project No. 025/2015/AMJ,FDCT,SAR Macau, China, and University of Macau Grant Nos: MYRG2018-00237-RTO, CPG2019-00004-FST and SRG2018-00111-FST.

7. REFERENCES

[1] Dietmar Jannach and Malte Ludewig, "When recurrent neural networks meet the neighborhood for session-based recommendation," in *RecSys*. ACM, 2017, pp. 306–310.

[2] Dilruk Perera and Roger Zimmermann, "Exploring the use of time-dependent cross-network information for personalized recommendations," in *ACM MM*, 2017, pp. 1780–1788.

[3] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin, "Personalized key frame recommendation," in *SIGIR*. ACM, 2017, pp. 315–324.

[4] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin, "Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews," in *SIGKDD*. ACM, 2017, pp. 717–725.

[5] Junyu Gao, Tianzhu Zhang, and Changsheng Xu, "A unified personalized video recommendation via dynamic recurrent neural networks," in *ACM MM*, 2017, pp. 127–135.

[6] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li, "Contextual video recommendation by multimodal relevance and user feedback," *TOIS*, vol. 29, no. 2, pp. 10, 2011.

[7] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel, "Multimodal neural language models," in *ICML*, 2014, pp. 595–603.

[8] Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang, "Crowdsourced time-sync video tagging using temporal and personalized topic modeling," in *SIGKDD*. ACM, 2014, pp. 721–730.

[9] Ming He, Yong Ge, Le Wu, Enhong Chen, and Chang Tan, "Predicting the popularity of danmu-enabled videos: A multi-factor view," in *DASFAA*. Springer, 2016, pp. 351–366.

[10] Ofer Tchernichovski, Marissa King, Peter Brinkmann, Xanadu Halkias, Daniel Fimiarz, Laurent Mars, and Dalton Conley, "Tradeoff between distributed social learning and herding effect in online rating systems: Evidence from a real-world intervention," *SAGE Open*, vol. 7, no. 1, pp. 2158244017691078, 2017.

[11] Qing Ping, "Video recommendation using crowdsourced time-sync comments," in *RecSys*, 2018, pp. 568–572.

[12] Wenmian Yang, Na Ruan, Wenyuan Gao, Kun Wang, Wensheng Ran, and Weijia Jia, "Crowdsourced time-sync video tagging using semantic association graph," in *ICME*. IEEE, 2017, pp. 547–552.

[13] Guangyi Lv, Tong Xu, Enhong Chen, Qi Liu, and Yi Zheng, "Reading the videos: Temporal labeling for crowdsourced time-sync videos based on semantic embedding.," in *AAAI*, 2016, pp. 3000–3006.

[14] Zhenyu Liao, Yikun Xian, Xiao Yang, Qinpei Zhao, Chenxi Zhang, and Jiangfeng Li, "Tscset: A crowdsourced time-sync comment dataset for exploration of user experience improvement," in *IUI*. ACM, 2018, pp. 641–652.

[15] Julian McAuley and Jure Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *RecSys*. ACM, 2013, pp. 165–172.

[16] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang, "Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars)," in *SIGKDD*. ACM, 2014, pp. 193–202.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[18] Ruining He and Julian McAuley, "Vbpr: Visual bayesian personalized ranking from implicit feedback.," in *AAAI*, 2016, pp. 144–150.

[19] Chaoqun Duan, Lei Cui, Xinchi Chen, Furu Wei, Conghui Zhu, and Tiejun Zhao, "Attention-fused deep matching network for natural language inference.," in *IJCAI*, 2018, pp. 4033–4040.

[20] Wei Qian, Cong Fu, Yu Zhu, Deng Cai, and Xiaofei He, "Translating embeddings for knowledge graph completion with relation attention mechanism.," in *IJCAI*, 2018, pp. 4286–4292.

[21] Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia, "Neural relation extraction via inner-sentence noise reduction and transfer learning," in *EMNLP*, 2018, pp. 2195–2204.

[22] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua, "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention," in *SIGIR*. ACM, 2017, pp. 335–344.

[23] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *RecSys*. ACM, 2017, pp. 297–305.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[25] Mike Schuster and Kuldip K Paliwal, "Bidirectional recurrent neural networks," *TIP*, vol. 45, no. 11, pp. 2673–2681, 1997.

[26] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky, "The Stanford CoreNLP natural language processing toolkit," in *ACL*, 2014, pp. 55–60.

[27] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.