Context-Dependent Propagating-Based Video Recommendation in Multimodal Heterogeneous Information Networks

Lei Sang, Min Xu, *Member, IEEE*, Shengsheng Qian, Matt Martin, Peter Li, and Xindong Wu, *Fellow, IEEE*

***Abstract*—With the emergence of online social networks (OSNs), video recommendation has come to play a crucial role in mitigating the semantic gap between users and videos. Conventional approaches to video recommendation primarily focus on exploiting content features or simple user-video interactions to model the users' preferences. Although these methods have achieved promising results, they fail to model the complex video context interdependency, which is obscure/hidden in heterogeneous auxiliary data from OSNs. In this paper, we study the problem of video recommendation in Heterogeneous Information Networks (HINs) due to its excellence in characterizing heterogeneous and complex context information. We propose a Context-Dependent Propagating Recommendation network (CDPRec) to obtain accurate video embedding and capture global context cues among videos in HINs. The CDPRec can iteratively propagate the contexts of a video along links in a graph-structured HIN and explore multiple types of dependencies among the surrounding video nodes. Then, each video is represented as the composition of the multimodal content feature and global dependency structure information using an attention network. The learned video embedding with sequential based recommendation are jointly optimized for the final rating prediction. Experimental results on real-world YouTube video recommendation scenarios demonstrate the effectiveness of the proposed methods compared with strong baselines.**

***Index Terms*—Video recommendation, context-dependent propagating, Heterogeneous Information Network (HIN), Network embedding.**

I. Introduction

As THE amount of data from different platforms grows at an unprecedented rate, video recommendation is becoming increasingly crucial to help users discover personalized content from this ever-growing corpus of data [1]. For example, 500 hours of videos are uploaded to YouTube every minute [2]. It is thus impossible for the users to watch all available videos to identify their interested ones. As a result, there is an enormous demand for video recommendation to provide relevant information tailored to the users' interests or preferences. Furthermore, the rapid emergence of user-generated content (UGC) and online social networks (OSNs) has provided a new rich library of social media content [3]–[5], which introduces new challenges and opportunities to the recommendation process. In this paper, we aim to design an effective and robust video recommendation method for OSNs.

Early works on video recommendation mainly include content-based methods [6] and collaborative filtering based methods [7]. As shown in Fig. 1(a), content-based methods recommend a user to watch videos similar to those have been watched before. As shown in Fig. 1(b), collaborative filtering based methods predict the interests of a user based on the user-video interactions, in the form of a graph/network structure. Content-based methods only calculate the video relevance using multimedia data analysis, which cannot directly reflect the users' general interests. Collaborative filtering based methods suffer from the cold-start problem and the sparsity problem of the user-video interaction matrix [8]. Currently, various types of auxiliary data, such as the user social relationships and video attributes (e.g., category and tag), are becoming increasingly available in OSNs [9]. Given the above limitations of the pure content-based and collaborative filtering based methods, many methods have been further proposed to leverage this auxiliary information to improve the recommendation performance [5], [10]–[12]. These methods can be considered hybrid methods for recommendation systems. However, due to the heterogeneity

Manuscript received January 30, 2019; revised October 17, 2019 and March 23, 2020; accepted June 16, 2020. Date of publication July 8, 2020; date of current version June 25, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1000901, in part by the China Scholarship Council (CSC), in part by the National Natural Science Foundation of China under Grant 91746209, and in part by the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education of China under Grant IRT17R32. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet. *(Corresponding author: Min Xu.)*

Lei Sang is with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, China, and also with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: lei.sang@student.uts.edu.au).

Min Xu is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia (e-mail: Min.Xu@uts.edu.au).

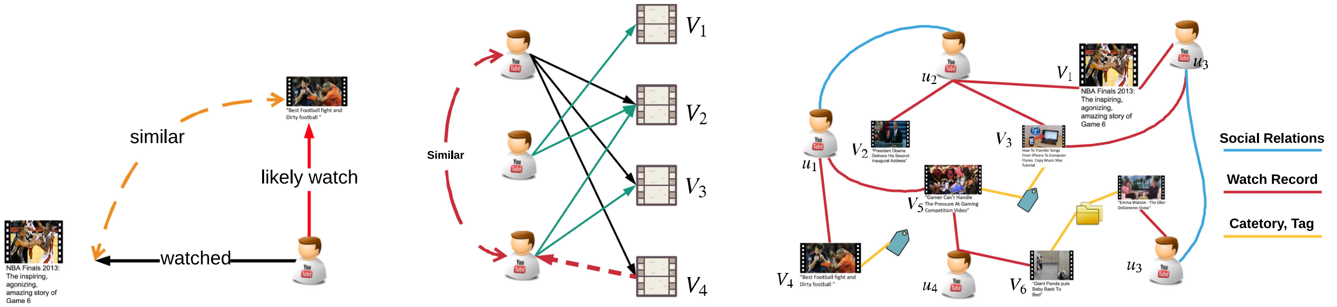
Shengsheng Qian is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shengsheng.qian@nlpr.ia.ac.cn).

Matt Martin and Peter Li are with the INTERACT Technology, Sydney, NSW 2000, Australia (e-mail: matt@interact.technology; peter@interact.technology).

Xindong Wu is with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education, Hefei University of Technology, Hefei 230009, China, and also with the Mininglamp Academy of Sciences, Mininglamp Technology, Beijing 100084, China (e-mail: xwu@hfut.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.3007330


similar
likely watch
watched
"Best Football fight and Dirty football "
NBA Finals 2013: The inspiring, agonizing, amazing story of Game 6
Similar
V_1
V_2
V_3
V_4
u_1
u_2
u_3
u_4
V_5
V_6
"President Obama Delivers His Second Inaugural Address"
How To Transfer Songs From iPhone To Computer iTunes Copy Music Mac Tutorial
"Gamer Can't Handle The Pressure At Gaming Competition Video"
"Emma Watson - The Ellen DeGeneres Show"
"Giant Panda puts Baby Back To Bed"
Social Relations
Watch Record
Catetory, Tag

(a) Content-based recommendation

(b) Collaborative filtering based recommendation

(c) General idea to adopt a multimodal HIN for recommendation

Fig. 1. (a) An example of content-based recommendation. Only the video content feature is used to model user preference. (b) Collaborative filtering uses simple user-video interaction for video recommendation. (c) Our general idea of introducing a multimodal HIN to model user historical information. We try to encode global video context and multimodal video content feature into the comprehensive video embedding for recommendation purposes.

and complexity of social data, it remains challenging for recommendation systems to effectively utilize such auxiliary context information.

Heterogeneous Information Networks (HINs) consist of various patterns of connections among different types of objects, which can reveal hidden interdependency among them [13], [14]. In this paper, we study the problem of video recommendation in HINs considering their flexibility in exploiting rich relations in various types of auxiliary data. An illustration of a video recommendation problem in an HIN can be found in Fig. 1(c). Our HIN includes all types of objects (e.g., videos, users and attribute information) and demonstrates different types of relations among objects, such as social relationships and user-attribute informaiton. These complex relations captured in the HINs provide us with a deep and latent perspective, from which to analyse the video interdependency relation among videos. In fact, all these relations can contribute to the recommendation task in an explicit or implicit manner. For example, the relation of two videos can be inferred by 1) "video-user-user-video" and 2) "video-category-video". These meta-paths capture the semantic relations of 1) being watched by friends or 2) belonging to the same genre. Hence, we can find potentially interesting videos based on the intuition that the paths connecting two videos represent the video relations of different semantics. Such an intuition facilitates the inference of user preferences based on video similarity to generate effective recommendations. This concept is known as HIN-based recommendation.

In the context of OSNs, video recommendation via HINs is still an open and challenging problem for several reasons. First, there is a general lack of consideration of how videos propagate through social connections, which can be used to model the complex and deep contextual interdependencies among videos. Nearly all existing models have leveraged the HIN-based structure in a simple manner by considering the local neighbours of each video [15], [16]. Social networks provide a fundamental medium for the diffusion and spread of information to neighbour videos via users' social connection and subsequently to the video neighbours, which forms a hierarchical interdependency among videos [17]. Instead of employing the one-hop local social network structure, how should we capture the influence-propagating process among videos for better social recommendation performance is an urgent issue that must be investigated.

Second, most of the existing HIN-based recommendations use the path-based similarity without considering the multi-modal video content feature of nodes (such as video title and description). Limited efforts have been devoted to analysing both graph structure and node feature in a unified manner to model the user preference [14]. As shown in Fig. 1(c), according to the meta-paths "video-user-user-video" (e.g., $V_1 - u_2 - u_1 - V_4$ and $V_1 - u_2 - u_1 - V_5$), video V_1 may have identical similarity to videos V_4 and V_5, since they have been watched by the same friends. However, videos V_4 and V_5 have different styles due to their different text and visual features. This example highlights that the same meta-path that connects a different video pair often carry relations of different semantics. In fact, V_1 and V_4 should be more similar, since they both belong to the category "ball game". The conventional meta-path does not allow a node to have content features [17], so it cannot reveal this subtle difference. To fully exploit paths in HINs for a recommendation, one must capture the semantics of different paths and their distinctive content feature in describing user preferences towards videos.

To exploit HINs for recommendation and overcome the above limitations, we propose a Context-Dependent Propagating Recommendation network, or CDPRec for brevity. Our goal is to explore both multimodal content features and context structure information from a multi-modal HIN to generate high-quality video embeddings for recommendation. The network embedding projects the nodes of a network[1] into a vector space. Compared to the pure meta-path based similarity, the network embedding approach is more resistant to noisy and sparse data. For a graph structure of the multi-modal HIN, we first extract a homogeneous video graph from the original HIN with a path-based random walk, which can be easier to handle while maintaining their original interdependency structure. For video content of the multi-modal HIN, we learn the shared video representation using a multimodal fusion layer that combines both visual and

[1] The terms 'graph' and 'network' have identical meaning here.

textual content. The proposed CDPRec has two main operations by which it learns the final video embedding from both graph structure and multimodal video features: 1) Propagation: The CDPRec can iteratively propagate a video's contexts along links in a graph-structured HIN and explore multiple types of dependencies among the surrounding video nodes. The propagating operation stimulates the diffusion of users' preference in a social network structure to discover the video in which she/he has the most hierarchical potential to be interested; 2) Aggregation: We collectively aggregate multimodal features of potential interests for each video node with an attention network, followed by a linear transformation to generate a new representation for a target video. Finally, an RNN-based sequential recommendation layer is seamlessly integrated with the network such that the CDPRec can be trained in an end-to-end manner.

The main contributions of this paper are as follows:

1) We are the first to propose a multimodal HIN embedding method to simultaneously uncover both multimodal video content and complex structural information of HINs. Moreover, we propose a meta-path-based random walk strategy to extract a homogeneous video graph and learn the shared video content feature using a multimodal fusion method.
2) We propose a Context-Dependent Propagating Recommendation network, called CDPRec for short. CDPRec is an embedding-based method that automatically discovers the hierarchically ranked potential interesting videos of users along links by iteratively propagating video context in the graph and aggregating them for video embedding.
3) We conduct experiments on a real-world YouTube dataset, the results of which demonstrate the efficacy of the CDPRec over strong baselines. In particular, the proposed model performs well in the cold-start scenario and has potentially good interpretability for the recommendation results.

II. Preliminary

The popularity of online social networks has resulted in a large amount of information being made available, such as social relationship and object attribute information (e.g., category and tag). Thus, it is straightforward to find the social network of users to alleviate the sparsity problem of simple user-video interaction. Furthermore, the emergence of tags and categories has allowed video uploaders to conveniently organize their videos, while also allowing the users to find interesting videos according to a given tag or category. Taking YouTube as an example, users may wish to seek out a group of interesting guitar artists by querying the "guitar" tag. We can use this rich auxiliary information to improve the recommendations.

HIN is a powerful tool to organize the above auxiliary information. Users and videos with few prior interactions can be linked together through different types of paths. We set up our recommendation system in the framework of the HIN, which can be defined as follows:

Definition 1. Heterogeneous Information Network (HIN) [18]: A HIN is denoted as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which consists of a node-type mapping function $\phi : \mathcal{V} \rightarrow \mathcal{A}$ and a link-type mapping function $\varphi : \mathcal{E} \rightarrow \mathcal{R}$. $\mathcal{A}$ and $\mathcal{R}$ are the sets of pre-defined node and link types respectively, and satisfy $|\mathcal{A}| + |\mathcal{R}| > 2$.

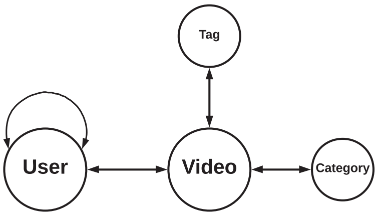


Fig. 2. Network schemas of a heterogeneous information network used for our social video recommendation scenario.

Generally, the HIN does not consider the content features on each node. However, real-world video recommendation contains rich multimodal content information on video nodes. Thus, we denote the HINs with multimodal video content features as *multimodal HINs*.

In HIN, the **network schema** is introduced to illustrate the high-level topology of a HIN, which describes the node types and their interaction relations.

Example 1: As shown in Fig. 2, we present the meta schema of our video recommendation system, which consists of various types of objects (e.g., User (U), Video (V), Tag (T) and Category (C)) and their semantic relations (e.g., 'friends' relation among users, 'watching' relation among users and videos, and 'attribute' relation among videos and tags/categories).

In HIN, nodes can be connected via multiple types of semantic paths, which are defined as **meta-paths**.

Definition 2. Meta-Path: Given a HIN, meta-path ρ is defined in the form of $\mathcal{A}_1 \xrightarrow{\mathcal{R}_1} \mathcal{A}_2 \xrightarrow{\mathcal{R}_2} \cdots \xrightarrow{\mathcal{R}_l} \mathcal{A}_{l+1}$ (abbreviated as $\mathcal{A}_1\mathcal{A}_2 \cdots \mathcal{A}_{l+1}$), which denotes a composite relation $\mathcal{R}_1 \circ \mathcal{R}_2 \circ \cdots \mathcal{R}_l$ between nodes $\mathcal{A}_1$ and $\mathcal{A}_{l+1}$, where $\circ$ is a composition operator on the relations.

Example 2: There exist multiple specific paths under the meta-path, which is called a path instance denoted by ρ. We can connect two videos with different meta-paths, e.g., "$Video - Category - Video$" (VCV) and "$Video - User - User - Video$" (VUUV). Commonly, different meta-paths can reveal the different interdependency information of two videos. Path "VCV" path illustrates that two videos belong to the same category, while path "VUUV" illustrates two videos watched by two friends.

Most existing HIN-based recommendations rely on meta-path based similarities for recommendation [13], [19], which may result in complex interdependencies among videos in the HINs being missed. In the following part, we propose a new graph embedding-based method for this task, which can mine both complex graph structure and multimodal node feature information hidden in HIN. Notations used throughout the article can be found in Table I.

III. Method

In this section, we present the proposed Context-Dependent Propagating network for Recommendation (CDPRec). Our task

TABLE I
NOTATIONS AND EXPLANATIONS

Notations	Description
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	a heterogeneous information network (HIN)
$\mathcal{V}$	HIN node set
$\mathcal{E}$	HIN link set
$\mathcal{A}$	HIN node type set
$\mathcal{R}$	HIN link type set
ρ	a meta-path
$G = (V, A, X)$	a homogeneous video graph
V	video node set
V_i	ith video in V
A	adjacency matrix
$e_{ij} \in A$	edge weight between video V_i and V_j
N	number of videos
$N(i)$	neighbor videos for video V_i
T_v	transformation matrix w.r.t. the visual content
T_t	transformation matrix w.r.t. the textual content
v_e	visual content embedding
t_e	textual content embedding
X	multimodal video feature
$x_i \in X$	embedding of video V_i
$N \times F$	dimension of X
X^L	output video embedding in layer L
$N \times F^L$	dimension of X^L
h_t^u	user preference

is to leverage both multimodal content features and the complex graph structure of the HINs to generate high-quality video embeddings. Then, these embeddings are used to generate recommender system candidates by computing the predicted probability. The extracted video embeddings are expected to fully capture the semantic meanings of video and complex video interdependency encoded in the HIN.

A. Framework

The overall framework of the proposed CDPRec is illustrated in Fig. 3, which consists of four major components: 1) For the graph structure, we extract the homogeneous video graph from the HINs via a meta-path-based random walk (Fig. 3(a)). This homogeneous video graph is easier to handle while still preserving its original contextual structure; 2) For the multimodal video node feature, we learn the shared video representation using a multimodal fusion layer that combines both visual and textual content (Fig. 3(b)). 3) To learn the final video embedding from the above graph structure and multi-modal video node feature, several propagating layers are used to iteratively extend the user's interests along graph links to discover his potential hierarchical interests and aggregate them for embedding learning purposes (Fig. 3(c)). 4) Finally, we extend the classical RNN-based sequential recommendation by incorporating the learned video embedding to compute the predicted probability for potential videos (Fig. 3(d)). In the following subsections, we describe each component of the CDPRec in detail.

B. Construction of Video Graph From HIN

For the convenience of network embedding learning for video, our first building block is to extract a homogeneous video graph from the HIN. Existing network embedding methods primarily focus on homogeneous networks where the nodes are of the same type, and they cannot be directly applied to heterogeneous networks. Since our main goal is to learn effective video embedding for a recommendation purpose, nodes of other types other than videos are of less interest in our task. Hence, we use a set of meta-path over the original heterogeneous graph to transform the original HIN into a homogeneous video graph, which is easier to handle while still preserving the original contextual structure among videos.

The key to generating a meaningful homogeneous video graph involves designing an effective method to capture the complex semantics reflected in HINs [14]. In general, three steps are required to generate the homogeneous video graph: 1) Inspired by [14], a meta-path-based random walk is adopted to generate a meaningful context sequence; 2) We filter out unrelated nodes to generate a pure video sequence; 3) Finally, we construct the video graph by considering the co-occurrence of video in sequence and drawing edges among them. In the following, we describe each step in detail.

In this paper, we design meta-path-based random walks to generate paths that are able to capture both the semantic and original structural correlations between different videos, facilitating the transformation of heterogeneous network structures into a homogeneous video graph. Given a HIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a meta-path $\rho : A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots A_t \xrightarrow{R_t} A_{t+1} \cdots \xrightarrow{R_l} A_{l+1}$, the transition probability of next walk is:

$$P(n_{t+1}|n_t, \rho) = \begin{cases} \frac{1}{|N^{(A_{t+1})}(v)|} & (n_t, n_{t+1}) \in \mathcal{E}, \phi(n_{t+1}) = A_{t+1} \\ 0 & (n_t, n_{t+1}) \in \mathcal{E}, \phi(n_{t+1}) \neq A_{t+1} \\ 0 & (n_t, n_{t+1}) \notin \mathcal{E} \end{cases} \quad (1)$$

where $n_t \in A_t$ is the t-th node in the random walk sequence, and $N^{(A_{t+1})}(n_t)$ denotes the neighbour node set for n_t with the type of A_{t+1}.

A meta-path-based random walk will iteratively follow the predefined path pattern, which ensures that different types of semantic relationships among video nodes can be properly preserved. For example, in the traditional random walk procedure in Fig. 3(a), the next step of a walker on node V_1 can be all types of nodes surrounding it: U_2 and T_1. However, under the meta-path **(V-U-U-V)**, the walker is biased to choose U_2 in the next step. Repetitively following the semantics of this path until it reaches the pre-defined length, we can sample a sequence "$V_1 \rightarrow U_2 \rightarrow U_4 \rightarrow V_5 \rightarrow U_3 \rightarrow U_1 \rightarrow V_2$". The meta-path-based random walk strategy ensures that the semantic relationships among different types of nodes can be properly preserved. A fundamental property of a random walk is that in the limit, the long-term average probability of being at a particular node is independent of the start node. Thus, the random walk can initiate from different seed nodes. In this paper, we use four types of video-related meta-path to generate a node sequence, where the starting type is V. (1) Two video co-occurrences appear in a user's watch list **(V-U-V)**. (2) Two videos have been watched by friends **(V-U-U-V)**. (3) Two videos have the same category **(V-C-V)**. (4) Two videos have the same tag **(V-T-V)**.

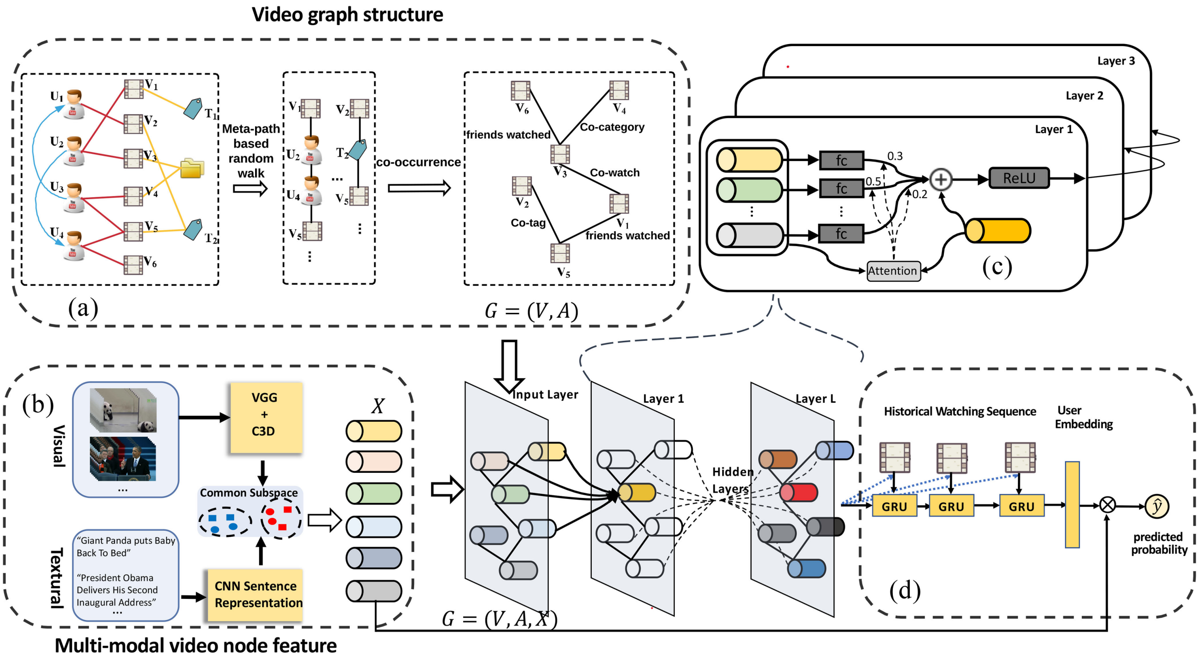


Fig. 3. The overall framework of our method. (a) For a graph structure, we use a meta-path based random walker to extract the homogeneous video graph $G = (V, A)$ from the HIN while preserving the original contextual structure among the videos. (b) For a node feature, both visual features (VGG and C3D) and textual features (CNN sentence model) are extracted to jointly learn the multi-modal video content feature X. (c) A context-dependent propagating network is exploited to iteratively extend a user's potential interests along the links and aggregate the video content of the potential interest node into a single embedding, which can naturally incorporate both complex graph structure and multimodal video content information. By stacking multiple layers, the final hidden representation of each video node receives dependency messages across further context steps. (d) Afterwards, the learned new video embeddings are fed into a GRU encoder for user preference modelling. Finally, the predicted probability is measured by considering the relevance between the user preference and a new candidate video. To train our model, the video node feature extraction (part a) and video embedding learning (part b) are first independently trained. The obtained node feature and node structure are fed into the GCN (part c). The GCN and sequential recommendation via the RNN are jointly trained for a final recommendation prediction.

Since our main focus is to learn a representation for videos, we filter out the nodes of other types and extract the video sequence based on the node sequence generated above. Thus, the final random walk sequence will contain only video nodes. To construct the homogeneous video graph, we consider the co-occurrence of videos in the sampled paths by setting a slide window interval [20]. Two videos will be connected by an edge if they occur in the same window interval. Then, we embed the video node over this homogeneous video graph. The most significant advantage of this approach is that we can relax the challenge of handling complex multiple node types in the HIN. In addition, this homogeneous video graph can maintain the original relation structure information among videos in the HIN because the links among homogeneous video nodes are essentially extracted from the heterogeneous local neighbour videos. Taking the meta-path V-U-U-V as an example, we can sample a sequence "$V_1 \rightarrow U_2 \rightarrow U_4 \rightarrow V_5 \rightarrow U_3 \rightarrow U_1 \rightarrow V_2$" according to Eq. 1. After the sequence is constructed via a pre-defined meta-path, we discard the nodes with different types from the video. Hence, we obtain the homogeneous video node sequence "$V_1 \rightarrow V_5 \rightarrow V_2$".

In addition to the node structure, we assign a weight to each edge e_{ij} based on the total number of occurrences of the two connected videos. More specifically, the weight of the edge is equal to the frequency of video i transitioning to video j in the entire behaviour, category and interaction history of the user. Thus, the constructed video graph can represent the similarity among different videos based on all user behaviours. The path length for the random walk and the fixed slide window size is empirically set. Finally, the homogeneous video graph structure is defined as $G = (V, A)$, where V is the set of video nodes, and A is the adjacency matrix. Each entry $e_{ij} \in A$ is the weight between video nodes V_i and V_j.

C. *Multimodal Video Feature Representation*

In addition to the graph structure, we learn the high-level semantic representation of video content. Commonly used content features for video recommendation are textual features [21], [22], such as video title and description. However, recent user-generated textual descriptions for video may be incomplete and noisy, so the existing approaches fail to generate precise content-based video recommendations in most cases. Therefore, some methods use multimodal video content (e.g., visual and texture content) for video recommendation [23], [24]. These methods mainly use hand-crafted weights to fuse information across different modalities, which may not be able to handle large-scale video data in recommendation systems [25]–[27].

The multimodal video feature representation in this paper aims to learn a joint video representation from textual and visual content by mapping them both into a shared space. Thus, the video content in the HINs can be comprehensively represented, which makes it easy to reveal the user preferences. For the visual content representation **v**, we extract the $fc6$ layer in VGG19 [28] and the last fully connected layer from C3D [29] for the frame and clip representations, respectively. The sampled frames and clips are processed by mean pooling and concatenated together to generate a single feature vector. The textual content representation **t** is extracted from the last layer of a pre-train convolutional text classification model [30], which takes Glove [31] as the word representation input.

We assume that there is a common space across visual and textual modalities, where the similarity between visual and textual contents can be measured. To map these two modalities into a common space, we design two linear mapping functions:

$$\mathbf{v}_e = \mathbf{T}_v \mathbf{v} \quad \text{and} \quad \mathbf{t}_e = \mathbf{T}_t \mathbf{t} \tag{2}$$

where $\mathbf{T}_v \in \mathbb{R}^{F \times D_v}$ and $\mathbf{T}_t \in \mathbb{R}^{F \times D_t}$ are the trainable transformation matrices that map the video content and semantic sentences into the common embedding. D_t, D_v, and F are the dimensions of the textual content, visual content, and common multi-modal video feature, respectively. Then, we obtain the shared video representation by concatenating them together: $\mathrm{X} = \mathbf{v}_e \oplus \mathbf{t}_e$.

As shown in Fig. 3(b), we optimize the common embedding by minimizing a pairwise loss as proposed in [32]. We use $score(\mathbf{v}_e, \mathbf{t}_e)$ to measure the similarity score between a visual-text pair. Moreover, μ^+ and ρ^{2+} represent the mean and variance of the matched visual-text pair similarity distribution, while μ^- and ρ^{2-} denote the mean and variance of the non-matching pair similarity distribution. In this paper, we try to 1) maximize the mean similarity score between non-matching pairs while also 2) minimizing the mean distance and variance of two distributions between matching pairs, which is defined as:

$$Loss = (\rho^{2+} + \rho^{2-}) + \lambda \max(0, \Delta - (\mu^+ - \mu^-)) \tag{3}$$

where λ is a term that balances the importance of two terms, and Δ is the max margin between the mean of the matching and non-matching distance distributions. $\mu^+ = \sum_{e=1}^{Q_1} \frac{score(\mathbf{v}_e, \mathbf{t}_e)}{Q_1}$ and $\rho^{2+} = \sum_{e=1}^{Q_1} \frac{score(\mathbf{v}_e, \mathbf{t}_e) - \mu^+}{Q_1}$ when visual feature v_e and text feature t_e belong to the same video. $\mu^- = \sum_{e=1}^{Q_2} \frac{score(\mathbf{v}_e, \mathbf{t}_e)}{Q_2}$ and $\rho^{2-} = \sum_{e=1}^{Q_2} \frac{score(\mathbf{v}_e, \mathbf{t}_e) - \mu^-}{Q_2}$ when visual feature v_e and text feature t_e belong to different videos. We train Eq. (3) by the gradient descent with a mini-batch size of 200. Specifically, we sequentially select $Q_1 + Q_2 = 200$ visual-text pairs from the training set for each mini-batch in the experiments. We use random negative sampling to choose the non-matching pair and resample for each epoch.

D. Context-Dependent Propagating Network

After obtaining the video graph structure and multimodal feature vector of video nodes, we describe a deep neural network suitable for processing the node-feature-guided network embedding method to learn the video embedding. See Fig. 3(c) for an overview.

Let us revisit the homogeneous video graph $G = (V, A, X)$ constructed above, which provides the following information:

- A set of N videos that constitute the nodes of the homogeneous video graph. Each video node V_i is represented by vector $x_i \in X$ of the video node. Let X be a $N \times F$ matrix representation of the multimodal video content feature representation.
- A set of pairwise relations among all videos, which form the edges among the videos. The weight between videos V_i and V_j is represented by $e_{ij} \in A$. Similarly, let adjacency matrix A be a $N \times N$ matrix that represents the structure of the graph.

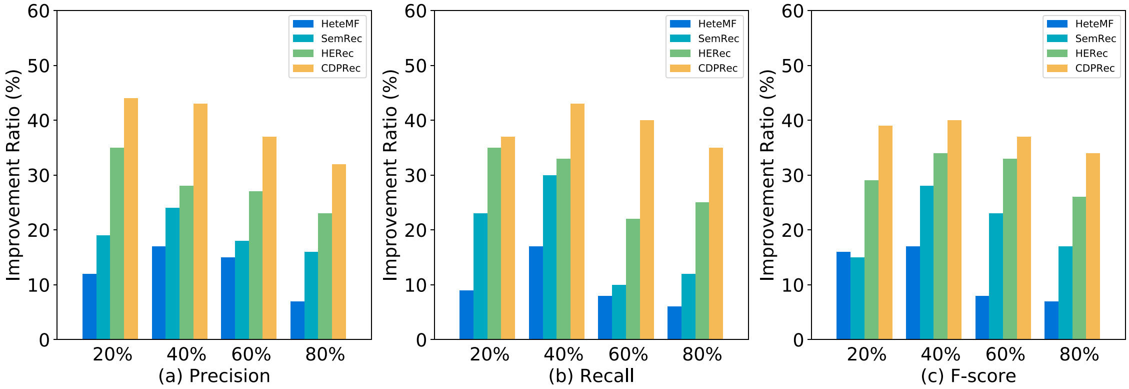
For the graph-structured data, we introduce a propagating model to improve the recommendation performance. Intuitively, if we can mine relevant videos based on the user preference in which a user was previously interested, then by following the discovered relevant videos, we can make video recommendations to this user accordingly. Relevant videos from the global neighbour context can be considered the natural extensions of a user's historical interests with respect to the HIN. For example, video V_3 has been co-watched with V_1, while video V_1 is further linked with a "friends watched" video V_5. The more steps we can go into the graph, the more context information we can obtain from increasingly further away on the graph.

Based on this observation, we design a context-dependent propagating layer for the video graph that borrows the key ideas of Graph Convolutional Networks (GCN) [33], [34] to simulate the video propagation along different meta-paths based social connections, which also enables us to generalise CNN to graphs. GCN is proposed to collectively propagate and aggregate information from the graph structure, which can thus model input and output consisting of the nodes' feature and their interdependency. Our proposed propagating layers have a feedforward architecture, and the input is node feature X with dimension and graph structure A. After passing through the first propagating layer, the output is a new node feature matrix X^1 with dimension $N \times F^1$ obtained by aggregating the node features of neighbourhood videos.

By stacking multiple layers, the final representation of each video node receives dependency messages across further context steps. One propagating layer encodes only context information about immediate neighbours and L layers are required to encode the L-order neighbourhoods (i.e., context interdependency among nodes at most L hops away). Each extra propagating layer extends the neighbour to a further-away context. Within each hidden layer, nonlinear activations can be applied to the node features X. After L layers, the final output node features $X^{(L)}$ can be considered an embedding of the graph nodes in an F^L-dimensional space. For simplicity, in one single propagating layer, let X and X' be the input and output node feature matrices with sizes $N \times F$ and $N \times F'$, respectively. This one-hop propagating process is defined as follows:

$$x'_i = \sigma\left(\sum_{j \in N(i)} e_{ij} x_j W + b\right) \tag{4}$$

where W is the trainable parameter with dimension $F' \times F$, and $N(i)$ is the neighbourhood of node i in the graph. $N(i)$ also includes v_i itself. The key idea behind the propagating layer is that the nodes progressively aggregate information from connected neighbours into each node's own representation; as this



(a) Precision (b) Recall (c) F-score

Fig. 4. Performance comparison of different methods for cold-start recommendation YouTube datasets. y-axis denotes the improvement ratio over CMF.

process iterates, we can mine the user's hierarchical potential interests from the global context of the graph. Thus, a target video node can incorporate the extended interest videos from friends co-watched, with the same tag or category.

1) Node Features Guided Attention: A critical concern about this aggregate mechanism is that different videos in a particular context may have different dependency weights for target nodes. We now introduce a form of attention into the aggregating process, which constitutes an essential part of the model. Graph Attention Network (GAT) [35] is a recently proposed method which computes the representations of each node in the graph. However, GAT only uses node features to decide the attention weights. The proposed attention function depends on both node features and edge features. The motivation is two-fold: (1) to identify the most relevant potential neighbour video to produce the recommendation; (2) to incorporate real-valued edge features that reveals the most frequently co-occurring videos. Practically speaking, we estimate the relevance weights of each possible neighbour video with the target video, which is also guided by the edge between them:

$$x_i' = \sigma\left(\sum_{j\in N(i)} \alpha(x_i, x_j, e_{ij}) x_j W + b\right) \tag{5}$$

where α is the so-called attention coefficient. α is a function of x_i, x_j and e_{ij}, which can indicate the importance of node j's features to node i. Moreover, our attention function is chosen to be the following:

$$\begin{aligned}\alpha(x_i, x_j, e_{ij}) &= \mathrm{softmax}\left(\mathrm{f}(\mathrm{x_i}, \mathrm{x_j})\mathrm{e_{ij}}\right)\\ &= \frac{\exp(f(x_i, x_j)e_{ij})}{\sum_{k\in N(i)} \exp(f(x_i, x_j)e_{ij})}\end{aligned} \tag{6}$$

where the attention function f is a single-layer feedforward neural network, parametrized by a weight vector $a \in \mathbb{R}^{2F'}$. Fully expanded out, the coefficients computed by the attention mechanism (illustrated by Fig. 4(c)) may then be expressed as:

$$\begin{aligned}&\alpha(x_i, x_j, e_{ij})\\ &= \frac{\exp(LeakyReLU\left(a^T[Wx_i||Wx_j]\right)e_{ij})}{\sum_{k\in N(i)} \exp(LeakyReLU\left(a^T[Wx_i||Wx_j]\right)e_{ij})}\end{aligned} \tag{7}$$

Algorithm 1: Propagating Layer With Single Head.

```
Input: Homogeneous video graph structure $G = (V, A)$;
  multimodal video feature $X$; depth $L$; weight matrices
  $W^l, \forall l \in \{1, \ldots, L\}$;; non-linearity $\sigma$;
Output: New video embedding $X^L$ for all video nodes;
$x_i^0 \leftarrow x_i, \forall v_i \in V$;
for $l = 1 \ldots L$ do
  for $v_i \in V$ do
    $\alpha \leftarrow softmax(ReLU((x_i^{l-1}||x_j^{l-1})e_{i,j})), V_j \in N(i)$;
$x_i^l \leftarrow \sigma(SUM(W^l \alpha x_j^{l-1}), V_j \in N(i))$;
  end for
end for
return $X^L \leftarrow x_i^L, \forall V_i \in V$
```

where $||$ is the concatenation operation and the LeakyReLU has negative input slope 0.2. The algorithm for the propagating layer update is given in Algorithm 1.

2) Multi-Head Attention: The single-head graph attention may suffer from stability problems; thus, we employ multi-head attention in this paper. The multi-head attention extracts multiple representations of the dependency among individual videos, which allows the model to jointly attend to information from different representation subspaces for different videos [36]. Moreover, with multiple convolutional layers, high-order relation representations can be extracted, which effectively capture the effect of other agents and greatly assist in co-operative decision-making. Specifically, K independent attention mechanisms execute the transformation of Eq. 4, after which their features are concatenated, resulting in the following output feature representation:

$$x_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N(i)} \alpha^k(x_i, x_j, e_{ij}) x_j W + b\right) \tag{8}$$

Unlike purely content or collaborative filtering based methods, the propagating layer method leverages both multimodal content information and graph structure. As a result, each video is modelled as a compact representation that considers the interdependency, heterogeneous and multimodal properties of the video recommendation.

E. Video Predicted Probability

A recurrent neural network (RNN) has been shown effective in capturing and characterizing the temporal dependency in sequence data. Following [37], to use the learned video embeddings for recommendation purposes, we feed each user's watch sequence into the RNN to learn the user preference from the watch history. Specifically, we adopt the gated recurrent unit (GRU) [38] network as the base sequential recommender in our work, since it is simpler and contains fewer parameters than the LSTM [39].

As shown in Fig. 3(c), we apply L-layers propagating over the video graph; then, each video node is represented by a new vector $x^L \in X^L$ based on equation 8; Given the historical watching sequence $\{i_1, \ldots, i_t\}$ of user u, our GRU-based recommender computes the current hidden state vector h_t^u conditioned on previous hidden state vector h_{t-1}^u as follows:

$$h_t^u = GRU(h_{t-1}^u, x_{i_t}^L; \Theta) \tag{9}$$

where GRU$(\cdot)$ is the GRU unit, $x_{i_t}^L$ is the embedding vector for item i_t, and Θ denotes all related parameters of the GRU networks. Thus, the predictor encodes the historical watching sequence of u into a hidden vector h_t^u, which models the sequential preference of u at t-th video in the user's interaction history. Hence, we call h_t^u the sequential user embedding of user u.

To generate the sequential recommendation, we rank a candidate video i by computing the recommendation score $\hat{y_i}$ according to:

$$\hat{y_i} = \sigma(h^{t^\top} x_i) \tag{10}$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is a sigmoid function. To train this model, we define the training loss as:

$$\mathcal{L} = \sum_{h_t^u, x^p, x^n} \max(0, -h_t^{u\top} x^p + h_t^{u\top} x^n + \Delta) \tag{11}$$

where h_t^u, x^p, x^n is a triplet pair. x^p is the positive video representation that denotes the next video that the user liked in groundtruth. We also sample negative videos x^n that the user does not interact with or has disliked. The basic idea is that we want to maximize the inner product of positive examples. Simultaneously, we want to ensure that the inner product of negative examples is smaller than that of the positive samples by a pre-defined margin Δ. In the training phase, the negative samples x^n are randomly selected from the training set and resampled each epoch.

To train our model, the video node feature extraction (part a) and video embedding learning (part b) are first independently trained. The obtained node feature and node structure are fed into the GCN (part c). The GCN and sequential recommendation via the RNN are jointly trained for the final recommendation prediction.

TABLE II
STATISTICS OF YOUTUBE DATASET

Users	Videos	Watch Records	Tags	Categories
6762	10894	169596	1000	33

IV. EXPERIMENTS

In this section, we conduct extensive experiments on real-world datasets to evaluate the proposed CDPRec method.

A. YouTube Video Dataset

To collect datasets containing both video and heterogeneous auxiliary information, we began with the Google+[2] site, which encourages its users to share their user accounts on other social networks. For simplicity, we adopted the user relationships from a cross-network dataset [1]. This dataset contains user account linkage between YouTube and Twitter and includes 143,259 Google+ users, among which 11,850 users provided both their Twitter and YouTube accounts. For each user relation, we downloaded the user's follower set from Twitter. For each user on YouTube, we further collected his/her watch-list over a one-year span of time from June 2013 to June 2014. However, the obtained data face data imbalance due to long-tail problems. Most users only watched a small proportion of the videos, and most of the videos were only watched a few times. If we were to train our model on such imbalanced datasets, the model would be dominated by observations with few active videos. To better evaluate the recommendation results, we filter the users by keeping the ones who watched over seven different videos on YouTube. Videos watched by fewer than three users are removed. Finally, we obtain 6,762 users and 10,894 videos in total for our experiment evaluation. For each video, we downloaded related information such as video categories, tags, titles, and descriptions via YouTube-dl.[3] We only keep the top 1,000 frequent tags in our experiment to avoid noise. Table II summarizes the key statistics of the YouTube data.

B. Evaluation Metrics

Within the experimental dataset, we train our proposed model utilizing 80% of the users, and both validation and test sets contain 10% of users to evaluate the recommendation results. We further design two different experimental modes: the short-range mode and the long-range mode. In the short-range mode, we use the first three months watch-list to model the user preference and keep the remaining nine months of watch-list as the objective for prediction. In the long-range mode, we use the first nine-month watch list to predict the remaining three-month watch-list. Following [40] and [19], we rank the videos by the predicted rating values and retrieve the top K videos, which is known as the top-K recommendation. The precision at rank K (Prec@K), recall at rank K (Recall@K) and F-score are employed as the evaluation metrics. We set $K = 5$ in our experiments.

[2] [Online]. Available: https://plus.google.com/

[3] [Online]. Available: http://rg3.github.io/YouTube-dl

To compare the performance of different algorithms, we add an evaluation metric AUC to evaluate the ranking performance of the recommendation results. AUC, which is the area under the ROC curve, is a widely used measure to evaluate the ranking performance:

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

where J denotes the positive samples set, and J' denotes the negative. $delta(p_{u,j} > p_{u,j'})$ is an indicator function that returns 1 if $(p_{u,j} > p_{u,j'})$ is true and 0 otherwise. $p_{u,j}$ is the predicted probability that user u may act on video V_i in the test set. Better ranking performance will yield a higher AUC value. An AUC larger than 0.5 indicates that the marker outperforms random guessing, and the best result is 1.

C. Baselines

We compare the proposed CDPRec with three types of representative recommendation methods: CF-based methods, which only leverages implicit user-video interaction information; content-based method, which analyses the video content to mine user preferences; HIN-based methods, which utilizes rich heterogeneous auxiliary information. To examine the effectiveness of incorporated multimodal content and the propagating mechanism, we prepare three variants of the CDPRec (CDPRec$_{dw}$, CDPRec$_{mg}$, CDPRec$_{propa}$). The comparison methods are given below:

- **LFM** [41]: A collaborative filtering based on the latent factor model that works by exploiting both explicit and implicit feedback by the users.
- **ItemKNN** [42]: The typical item-based top-N recommendation algorithm recommendation method, which recommends videos to a user using item-to-item similarities to compute the recommendations.
- **BPR-MF** [43]: A form of Bayesian personalized ranking based on the matrix factorisation method and a pairwise learning method for an item recommendation, which trains on pairs of positive observed interaction and negative unobserved counterparts of a user and takes the Matrix Factorization as the underlying predictor.
- **Bi-LSTM** [44]: Bi-LSTM (bi-irection long short-term memory) units are building units for the layer of a recurrent neural network (RNN), which makes sequential predictions. Here, we use the multi-modal video representation as features for the input.
- **Dynamic RNN** [37]: A dynamic RNN to model the dynamic interests of users over time by considering video semantic embedding and user relevance mining in a unified framework for video recommendation. Compared with the Bi-LSTM [42] method, [36] add additional user relationship constraint in the RNN. For a fair comparison, we did not include the additional twitter text for topic modelling.
- **CMF** [45]: A collective matrix factorization method, which extends the matrix factorization to multi-relational HINs by optimizing a joint objective over all relation types.
- **HeteMF** [16]: A matrix factorization-based recommendation method that uses meta-path-based item similarities for recommendation.
- **SemRec** [15]: A recommendation method utilized in weighted HINs that uses a weighted meta-path to obtain different preferences of users on the paths.
- **HERec** [14]: HERec is the state-of-art HIN-based ranking method that merges different meta-path guide node embeddings for item recommendation. For a fair comparison, we concatenate this video node embedding with the video content features to form the representation of a video.
- **CDPRec**$_{dw}$ A variant of the CDPRec, which can be viewed as a variant of the CDPRec and, incorporates the homogeneous network embedding method DeepWalk [46] and views all nodes in the HINs as being of the same type.
- **CDPRec**$_{mg}$ A variant of the CDPRec that incorporates the heterogeneous network embedding method of methpah2vec [47]. methpah2vec leverages a heterogeneous skip-gram model to exploit the meta-path heterogeneous neighbourhood of a node.
- **CDPRec**$_{prapa}$ Another variant of the CDPRec that ignores the multimodal video content feature. The video feature is obtained from the user/item interaction matrix with matrix factorisation model SVD.
- **CDPRec**: Our complete model.

We implement the CDPRec model using the Python library of TensorFlow. For our model, we randomly initialize the model parameters using the Gaussian distribution and optimize the model using adaptive moment estimation (Adam). During training, we apply L_2 regularization with $\lambda = 0.0005$. Furthermore, dropout with $p = 0.6$ is applied to the input of each propagating layer and the normalized attention coefficients. For MF-based recommendation methods, we follow the optimal configuration and architecture reported in [45]. For the other comparison methods, we optimize their parameters using 30% training data as the validation set. All experiments are conducted on a machine with four GPUs (NVIDIA GTX-1080 * 4), one CPU (i7-5960X CPU @ 3.00 GHz) and 48 GB memory.

D. Experimental Results

Table III reports the results of our proposed model and baselines on the YouTube dataset. We can make the following observations from the experimental results:

1) In all cases, our proposed complete model of the CDPRec outperforms all compared baseline methods in both short and long modes. Moreover, most of the improvements achieved by our method are significant compared to the best baseline methods. We attribute the superiority of the CDPRec to the following three properties: (a) semantic embedding of a video is learned to map multimodal visual and text information into a common semantic space to effectively integrate the content and extract deep interdependency structure information; (b) our context-dependent propagating mechanism adopts a more comprehensive method of improving the aggregation of the information of different meta-paths to improve the

TABLE III
EXPERIMENTAL RESULTS ON YOUTUBE DATA. WE INDICATE THE DATA SOURCE FOR EACH MODEL. (1) "CF" DENOTES THE USER-ITEM INTERACTION FOR COLLABORATIVE FILTERING PURPOSE. (2) "CT" DENOTES THE CONTENT FEATURE OF THE VIDEO. (3) "HIN" DENOTES THE HETEROGENEOUS INFORMATION NETWORK THAT CONSISTS BY USERS AND ATTRIBUTE INFORMATION OF VIDEOS. (4) "SOCIAL" DENOTES THE SOCIAL RELATIONSHIP AMONG USERS

		Short model			Long model			AUC
Model	Data Source	Precision	Recall	F-score	Precision	Recall	F-score	
LFM	CF	0.0024	0.0019	0.0021	0.0032	0.0046	0.0038	0.6513
ItemKNN	CF	0.0163	0.0127	0.0142	0.0169	0.0255	0.0203	0.6942
BPR-MF	CF	0.0182	0.0121	0.0145	0.0174	0.0273	0.0213	0.7128
Bi-LSTM	CF+CT	0.0231	0.0143	0.0176	0.0196	0.0349	0.0251	0.7417
CMF	HIN	0.0296	0.0181	0.0225	0.0238	0.0367	0.0288	0.7824
HeteMF	HIN	0.0321	0.0234	0.0270	0.0258	0.0376	0.0306	0.7939
SemRec	HIN + Social	0.0347	0.0247	0.0288	0.0287	0.0390	0.0330	0.7965
Dynamic RNN	CF + CT + Social	0.0349	0.0257	0.0296	0.0293	0.0411	0.0342	0.8123
HERec	HIN + social	0.0361	0.0264	0.0305	0.0306	0.0425	0.0355	0.8174
CDPRec$_{dw}$	HIN + social	0.0292	0.0226	0.0254	0.0241	0.0373	0.0293	0.7815
CDPRec$_{mp}$	HIN + social	0.0352	0.0253	0.0294	0.0316	0.0419	0.0360	0.8023
CDPRec$_{prapa}$	HIN + social	0.0383	0.0287	0.0328	0.0315	0.0431	0.0363	0.8338
CDPRec	HIN + CT + social	**0.0401**	**0.0314**	**0.0352**	**0.0346**	**0.0478**	**0.0401**	**0.8615**

recommendation performance; (c) attention improves the representations for the meta-path-based local neighbour video in a mutually beneficial manner.

2) Among the baseline methods, HIN-based methods (CMF, HeteMF, SemRec and HERec) outperform CF methods (ItemKNN, LFM, BPR-FM) and the content-based method Bi-LSTM in most cases, which demonstrates the benefit of auxiliary heterogeneous information.
Compared with the Bi-LSTM (CF + CT) [44] method, the Dynamic RNN (CF + CT + Social) [37] adds additional users' common interest (user relationship) constraint in the RNN. It achieves better performance than Bi-LSTM.
It is noteworthy that the recently proposed HERec model works well among these baselines, since this method adopts a similar path guided random walk to generate a context sequence for user/item embeddings.

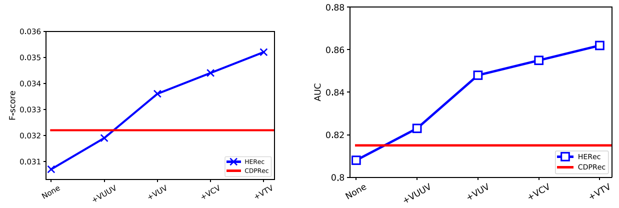
3) The proposed CDPRec (HIN + CT + Social) achieves better performance than Dynamic RNN (CF + CT + Social) [37]. It demonstrates that the powerful heterogeneity modelling ability of the HIN and our proposed context-dependent propagating method with attention provide additional ability to mine the complex context information and hidden interdependency in the HIN.

4) For the three embedding-based variants of the CDPRec, we notice that the order of overall performance is as follows: CDPRec $>$ CDPRec$_{propa}$ $>$ CDPRec$_{mp}$ $>$ CDPRec$_{dw}$. (a) CDPRec (with multimodal content feature) performs better than CDPRec$_{propa}$ (with spare user/video interaction feature), which indicates that the video content with visual and text information can provide discriminative features for video modelling. This additional information is particularly useful when analysing users with a small number of rating records. (b) CDPRec$_{propa}$ (with propagating-based embedding) outperforms CDPRec$_{mp}$ (with the meta-path-based similarity embedding). More importantly, both CDPRec$_{propa}$ and the strongest baseline HERec only leverage the graph structure information of the HIN and without considering the content feature. However, CDPRec$_{propa}$ outperforms HERec. The reason is that the propagating mechanism can effectively mine the high-order potential interests of the users, which may result in more useful context information being gained. (c) In addition, the improved performance of CDPRec$_{mp}$ relative to CDPRec$_{dw}$ confirms the benefit of the HIN-based embedding in modelling relation data of multiple types. (d) In summary, the CDPRec achieves the best performance under all circumstances. Based on these results, we argue that it is necessary to learn the deep independency structure and incorporate multi-modal video content features that can improve the recommendation performance.

E. Detailed Analysis of the Proposed Model

1) Cold-Start Recommendation: The "cold start" originates from the data sparsity problem. It is a real challenge for social recommendation to perform accurate recommendation using less training data. HIN is particularly useful for alleviating the cold-start problem in recommendation by leveraging the deep interdependency context information. In this part, we conduct experiments with different levels of cold-start data to validate whether the CDPRec can handle the cold-start problem. For comparison, we run baseline methods on the YouTube dataset with different sparsities. Here, we only use the HIN-based recommendation baselines, including CMF, HeteMF, SemRec and HERec. Following [19], we first randomly shuffle the dataset and split it into five equal groups. We set aside the last group as a hold-out or test data set; then, we vary the number of remaining groups from one to four, which correspond to 20%, 40%, 60%, and 80% of data being used as training sets.

For simplicity, we use the ratios of improvement w.r.t. CMF. The experimental results on the YouTube video dataset are presented in Fig. 4. Overall, all compared methods perform better than CMF. The proposed CDPRec method performs consistently better than all baselines; moreover, with fewer datasets being used, the improvement over CMF becomes more significant. The results show that various types of rich-context information in HINs are useful for providing better recommendation results,



(a) F-score (b) AUC

Fig. 5. Performance change of CDPRec when meta-paths are gradually added.

TABLE IV
THE RESULTS W.R.T. DIFFERENT LAYER NUMBER

Layer number L	1	2	3	4
Precision	0.0357	0.0374	0.0401	0.0386
Recall	0.0289	0.0291	0.0314	0.0307
F-score	0.0319	0.0327	0.0352	0.0341
AUC	0.8217	0.8464	0.8615	0.8357

and the proposed CDPRec approach can utilize this information in a more principled manner.

2) Impact of Different Meta-Paths: In this paper, we use different meta-path-based random walks to model local video contexts. To analyse the effect of different meta-paths on our final recommendation performance, we gradually add meta-paths into the CDPRec one by one and investigate the impact. For ease of analysis, we consider the HERec to be the basic reference. Fig. 5 shows, that the recommendation performance monotonically increases with the addition of more meta-paths. Meanwhile, meta-paths appear to have different effects on the recommendation performance. In particular, adding VUUV and VUV yields a significant performance boost, while the performance improvement resulting from adding VTV and VCV is relatively lower. The reason may be that user-generated video tags and categories will introduce more local context neighbours than social relation does, which can include noise or information that conflicts with existing neighbours when iteratively fusing them with the propagating mechanism.

3) Propagating Layer Number: We also vary the size of local graph context neighbourhoods by the maximal layer number L and observe the performance changes. The results are listed in Table IV, which clearly shows that when layer number L increases, the performance initially increases but eventually falls when L is too large (typically when $L = 3$). We attribute this phenomenon to the trade-off between useful information from the long-distance dependency and useless information from noise: 1) a too-small layer value L cannot adequately explore the deep interdependency and gain sufficient context information; and 2) a too-large value of L may lead to an overfitting problem when the number of parameters increases, while simultaneously introducing more undesired noise from the neighbours.

4) Density of Social Network: We explore how the density of social network affects the recommendation performance. The density of the social network is defined as $\frac{\#user\ relation}{\#user \times \#user}$, which is the ratio of filled entries in the social relation matrix to its size. To investigate the effect of the social network density to the recommendation, we gradually add social relation with different density levels, e.g. 10%–100% of original user links, by randomly eliminating non-zero entries of the social relation (to make it sparser). For better comparison, we also train the proposed model on HIN without social relation. The results are shown in Fig. 6. As we gradually increase the density of the social network, we can see that the recommendation performance increases in general with the density. It is worth noting that the performance is slightly jeopardised by lower social density, when keeping only 10% of the social relation. The probable reason may be that the model overfitting these small numbers of user relations and neglect other unseen relations. Overall, the performance of recommendation can benefit from auxiliary social network data, except for the extremely sparsity circumstance.

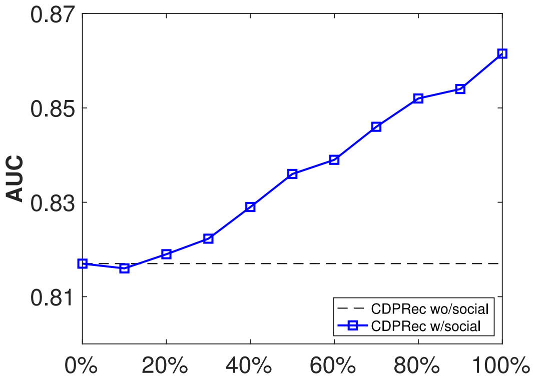


Fig. 6. Performance change of CDPRec with different density of social network(by keeping different proportions of social links).

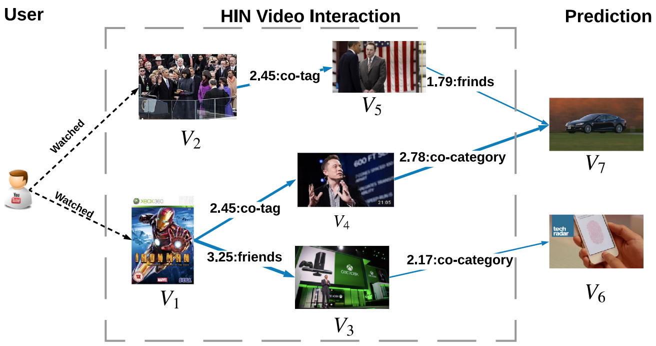
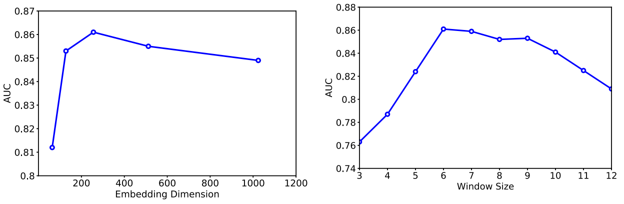


Fig. 7. Visualization of three paths with relevance probabilities for a user.

F. Case Study

Extra user-video connectivity information derived from the HIN endows the recommender systems the ability of reasoning on paths to infer the user preferences towards target items and generating reasonable explanations. To demonstrate this case, we show an example drawn from the CDPRec on a video recommendation task as shown in Fig. 7. We randomly sample a user with two historical watched videos $\{V_1, V_2\}$. Then, our system recommends video $\{V_6, V_7\}$ to this user. We extract all qualified paths that connect the user-item pair and present the subgraph in Fig. 6. For each multi-hop relevant video of the user, we calculate the (unnormalized) relevance probability between



(a) Video embedding dimension (b) Size of Co-occurrence window

Fig. 8. Parameter sensitivity of CDPRec.

the watched video $\{V_1, V_2\}$ and the predicted videos $\{V_6, V_7\}$. We have several observations.

The predicted video is connected to what the user has watched (e.g., V_1, V_2) by the shared video entities such co-friends viewed (V_3) and co-tag (V_4). This result shows that the CDPRec can extend the user interests along HIN paths. By analysing two paths for prediction V_7, we also find that different paths may describe the user-item connectivity from dissimilar angles, which can be treated as the evidence of why the item is suitable for the user. Our method provides a new viewpoint for the explainability by tracking the paths from a user's history to a new video with high relevance probability in the HIN, such as "$user \xrightarrow{watched} V_2 \xrightarrow{co-tag} V_5 \xrightarrow{friends-watched} V_7$" or "$user \xrightarrow{watched} V_1 \xrightarrow{co-tag} V_4 \xrightarrow{co-category} V_7$". Thus, recommendations tell the users what videos they may like while revealing the reason with which they may like them, which can help to improve the persuasiveness and user satisfaction of the recommender systems.

G. Parameter Sensitivity

In this section, we investigate how the recommendation performance varies with the hyper-parameters in 1) the dimension of final video embedding and 2) the length of the co-occurrence window when constructing the homogeneous video graph.

1) Dimension of Video Embedding: We vary the dimension of video embedding to 64, 128, 256, 512 and 1024. Fig. 7(a) shows the accuracy of the CDPRec over different video embedding dimensions. The accuracy shows an apparent increase at first because more bits can encode more useful information. However, the performance subsequently starts to slowly decrease when we continuously increase the number of dimensions. This result is intuitive, since a too-large dimension size requires more data to prevent an overfitting problem.

2) Length of Co-Occurrence Window: When constructing the homogeneous video graph from the HIN, the length of the context window controls the "density" of the video graph. As shown in Fig. 7(b), a larger window interval does not seem to help; on the contrary, a larger window corresponds to lower precision likely because a relation among videos that are further apart is not sufficiently strong to define a connection in the graph.

V. Related Work

In this section, we briefly review the related work including those in the areas of video recommendation, HIN-based recommendation and network embedding.

A. Video Recommendation

Recommendation is the most effective tool to alleviate the information overload problem on video-sharing websites [4], [23], [48]. Most related video recommendations may be roughly grouped under three categories: content collaboration [37], [49], collaborative filtering (CF) [7], [50] and hybrid methods [17], [51]. For content-based video recommendation, Mei *et al.* [49] present the contextual video recommender method VideoReach, which describes the relevance between two videos by fusing multimodal textual, visual, and aural information. Collaborative filtering represents users' preferences using a user-item matrix, which predicts a user's ratings of videos based on the preferences of similar users. Huang *et al.* [7] proposed a scalable matrix factorization-based collaborative filtering algorithm that combines the implicit feedback with an online updating strategy. However, collaborative-filtering-based methods may be unreliable when data are sparse or in cold-start user/item scenarios. Hybrid approaches combine content-based approaches and collaborative filtering with other auxiliary information to achieve better performance. Despite the promising results of these existing methods, effective video recommendation remains challenging, primarily due to the failure of the aforementioned methods to alleviate sparsity and the cold-start problem.

B. HIN-Based Recommendation

Modern social networks enable us to leverage additional information for recommendation, such as user/item attributes and social relationship. Accordingly, we can use the rich context provided by these heterogeneous data to gain better recommendation results. For example, Cui *et al.* [52] proposed a friends computing model to combine video content and social networks. In addition to social networks, Zhao *et al.* [53] develop a ranking-based video recommendation method that incorporates video meta-information such as title/tags and category. However, most of these existing methods independently address different heterogeneous information and fail to take advantage of the rich relation information among the objects.

As a result, the HIN-based recommendation has been proposed to model complex objects with different types and relation links. Many path-based similarity measurements [16], [18] have been developed to describe the relations among objects in HINs. Several methods have attempted to use these path-based similarities to improve the recommendation performance. The most similar work to ours is HERec [14], which proposed an HIN embedding approach by fusing different meta-path random-walk-based context information and without considering the vital node feature information. It is worth noting that the meta-path-based random walk is a common method to preserve the graph structure information, and is used in HINs [18], [47]. However, the objective of graph embedding in our task is to mine potential user interests through a propagating mechanism, which can naturally incorporate the deep interdependency structure and multimodal video feature information. The recently published SMR-MNRL [5] proposes a user-ranking model for the social-aware movie recommendation, which learns user embedding and movie embedding to predict the ranking score. The user embedding is learned from the HIN; however, the multimodal embedding of movies is merely learned from content information and do not consider the rich relation among them.

C. Network Embedding

Network embedding (NE) methods have shown outstanding performance on many tasks including node classification [46], [54], link prediction [55] and community detection [56], [57]. These methods aim to encode the network nodes into a low-dimensional vector space while preserving the network topology structure information. DeepWalk [46] uses skip-gram over a random-walk-based node sequence to learn the network representation. Recently, there have been several attempts to apply the notion of "graph convolutions" into network embedding, which has resulted in a new version of graph convolutions based on spectral graph theory. Typical examples include GCN [33], GraphSAGE [34], and the state-of-the-art model GAT [35].

The above methods are designed to embed a homogeneous graph with a single type of node. Several works have attempted to incorporate the structural information of an HIN into embeddings. Dong *et al.* [47] used meta-path-based random walks over HINs to construct the heterogeneous neighbourhood of a node; then, they fed these walk paths into a skip-gram model to generate the node embeddings. Fu *et al.* [58] proposed HIN2Vec, which adopts a logistic classification method to learn the node embeddings and meta-paths in an HIN. Although these methods can embed various heterogeneous networks, their representations ignore the deep interdependencies among the nodes and the rich content features of the nodes, so they may not be optimal for recommendation task. More recently, [59] modeled the attention coefficients of the HIN node in a more fine-grained manner by considering both path instance node and different meta-path representations.

VI. Conclusion

In this paper, we propose the CDPRec, a neural framework that incorporates HINs and multimodal video content into recommender systems in a natural and intuitive manner. The CDPRec overcomes the limitations of path-based similarity methods by introducing context-dependent propagation, which automatically propagates the potential preferences of the users and explores their hierarchical interdependencies in the HIN. The CDPRec encodes the global video context into comprehensive video embedding to facilitate the click-through rate prediction. Experimental results on real-world YouTube data demonstrate the significant superiority of our method over strong baselines.

References

[1] M. Yan, J. Sang, and C. Xu, "Mining cross-network association for youtube video promotion," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 557–566.

[2] [Online]. Available: "http://tubularinsights.com/hours-minute-uploaded-youtube/". Accessed on: Jul. 28, 2018.

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

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

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

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

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

[8] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender Systems Handbook*. Berlin, Germany: Springer, 2015, pp. 1–34.

[9] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1255–1267, Oct. 2013.

[10] B. Chen, J. Wang, Q. Huang, and T. Mei, "Personalized video recommendation through tripartite graph propagation," in *Proc. 20th ACM Int. Conf Multimedia*, 2012, pp. 1133–1136.

[11] X. Huang *et al.*, "Explainable interaction-driven user modeling over knowledge graph for sequential recommendation," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 548–556.

[12] X. Huang, S. Qian, Q. Fang, J. Sang, and C. Xu, "CSAN: Contextual self-attention network for user sequential recommendation," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 447–455.

[13] X. Yu *et al.*, "Personalized entity recommendation: A heterogeneous information network approach," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 283–292.

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

[15] C. Shi *et al.*, "Semantic path based personalized recommendation on weighted heterogeneous information networks," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 453–462.

[16] X. Yu, X. Ren, Q. Gu, Y. Sun, and J. Han, "Collaborative filtering with entity similarity regularization in heterogeneous information networks," in *Proc. IJCAI-HINA Workshop*, 2013.

[17] Q. Huang, B. Chen, J. Wang, and T. Mei, "Personalized video recommendation through graph propagation," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 10, no. 4, pp. 1–17, 2014.

[18] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "PathSim: Meta path-based top-k similarity search in heterogeneous information networks," in *Proc. VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

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

[20] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 404–411.

[21] J. Davidson *et al.*, "The youtube video recommendation system," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 293–296.

[22] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. 21rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1235–1244.

[23] P. Cui, Z. Wang, and Z. Su, "What videos are similar with you?: Learning a common attributed representation for video recommendation," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 597–606.

[24] X. Ma, H. Wang, H. Li, J. Liu, and H. Jiang, "Exploring sharing patterns for video recommendation on youtube-like social media," *Multimedia Syst.*, vol. 20, no. 6, pp. 675–691, 2014.

[25] Z. Ma *et al.*, "Discriminating joint feature analysis for multimedia data understanding," *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1662–1672, Dec. 2012.

[26] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 233–246, Feb. 2016.

[27] S. Qian, T. Zhang, R. Hong, and C. Xu, "Cross-domain collaborative learning in social multimedia," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 99–108.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2014.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4489–4497.

[30] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[32] B. Kumar, G. Carneiro, I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 5385–5394.

[33] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[34] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 1024–1034.

[35] P. Veličković *et al.*, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[36] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[37] J. Gao, T. Zhang, and C. Xu, "A unified personalized video recommendation via dynamic recurrent neural networks," in *Proc. ACM Multimedia Conf.*, 2017, pp. 127–135.

[38] K. Cho *et al.*, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process*, 2014, pp. 1724–1734.

[39] L. Wu, M. Xu, J. Wang, and S. Perry, "Recall what you see continually using gridlstm in image captioning," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 808–818, Mar. 2020.

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

[41] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 426–434.

[42] G. Karypis, "Evaluation of item-based top-n recommendation algorithms," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, 2001, pp. 247–254.

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

[44] Y. Zhang *et al.*, "Sequential click prediction for sponsored search with recurrent neural networks," in *Proc. Assoc. Advancement Artif. Intell.*, 2014, vol. 14, pp. 1369–1375.

[45] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 650–658.

[46] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 701–710.

[47] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for heterogeneous networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 135–144.

[48] R. Zhou, S. Khemmarat, and L. Gao, "The impact of youtube recommendation system on video views," in *Proc. 10th ACM SIGCOMM Conf. Internet Meas.*, 2010, pp. 404–410.

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

[50] S. Baluja *et al.*, "Video suggestion and discovery for youtube: Taking random walks through the view graph," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 895–904.

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

[52] L. Cui *et al.*, "A video recommendation algorithm based on the combination of video content and social network," *Concurrency Comput.: Pract. Experience*, vol. 29, no. 14, 2017, Paper e3900.

[53] X. Zhao *et al.*, "Integrating rich information for video recommendation with multi-task rank aggregation," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 1521–1524.

[54] L. Sang, M. Xu, S. Qian, and X. Wu, "AAANE: Attention-based adversarial autoencoder for multi-scale network embedding," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2019, pp. 3–14.

[55] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[56] X. Wang *et al.*, "Community preserving network embedding," in *Proc. Assoc. Advancement Artif. Intell.*, 2017, pp. 203–209.

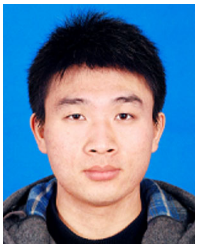
[57] L. Sang, M. Xu, S. Qian, and X. Wu, "Multi-modal multi-view bayesian semantic embedding for community question answering," *Neurocomputing*, vol. 334, pp. 44–58, 2019.

[58] T.-Y. Fu, W.-C. Lee, and Z. Lei, "HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2017, pp. 1797–1806.

[59] S. Zhou, J. Bu, X. Wang, J. Chen, and C. Wang, "HAHE: Hierarchical attentive heterogeneous information network embedding," 2019, *arXiv:1902.01475*.

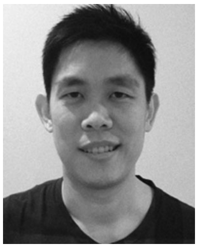
Lei Sang is currently working towards the Ph.D. degree with the School of Computer Science and Information Engineering, Hefei University of Technology, China, and also with the Faculty of Engineering and Information Technology, University of Technology, Sydney, Ultimo, NSW, Australia. His current research interests include natural language processing and recommender system.

Min Xu (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, in 2000, the M.S degree from the National University of Singapore, in 2004, and the Ph.D. degree from the University of Newcastle, Australia, in 2010. She is currently a Senior Lecturer with the University of Technology, Sydney. Her research interests include multimedia data analytics, pattern recognition and computer vision. She has published over 100 research papers in high quality international journals and conferences.



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

Matt Martin received the B.E. degree from The University of Technology Sydney, Sydney, Australia, in 1993. He is CEO of INTERACT Technology. He spent 20 years with pharmaceutical companies in a variety of selling and marketing roles. Matt build a digital communications platform nextINTERACT that could deliver rich content in a personalised format to doctors and other healthcare professionals.



Peter Li received the B.E. degree from The University of Sydney, Sydney, Australia, in 1994. He is current Chief Information Officer at INTERACT Technology. He is leading solution architect and mobile/desktop UI software engineer with 20 years I.T. experience, delivering solutions across the telco, digital media, IPTV industries, delivered mobile application for companies including Coca Cola, GrainCorp, MLC and Westpac.

Xindong Wu (Fellow, IEEE) received the Ph.D. degree in artificial intelligence from The University of Edinburgh, Edinburgh, U.K.. He is a professor with the Hefei University of Technology, China and the University of Louisiana at Lafayette, USA. His current research interests include data mining, knowledge-based systems, and Web information exploration. He is the Steering Committee chair of IEEE International Conference on Data Mining (ICDM). He is the editor-in-chief of Knowledge and Information Systems (KAIS) and ACM Transactions on Knowledge Discovery from Data (TKDD). He is a fellow of the AAAS.