What Videos Are Similar with You? Learning a Common Attributed Representation for Video Recommendation

Peng Cui, Zhiyu Wang and Zhou Su
Department of Computer Science and Technology, Tsinghua University, Beijing, China
cuip@tsinghua.edu.cn, zy-wang08@mails.tsinghua.edu.cn, suhmily@gmail.com

ABSTRACT

Although video recommender systems have become the predominant way for people to obtain video information, their performances are far from satisfactory in that (1) the recommended videos are often mismatched with the users' interests and (2) the recommendation results are, in most cases, hardly understandable for users and therefore cannot persuade them to engage. In this paper, we attempt to address the above problems in data representation level, and aim to learn a common attributed representation for users and videos in social media with good interpretability, stability and an appropriate level of granularity. The basic idea is to represent videos with users' social attributes, and represent users with content attributes of videos, such that both users and videos can be represented in a common space concatenated by social attributes and content attributes. The video recommendation problem can then be converted into a similarity matching problem in the common space. However, it is still challenging to balance the roles of social attributes and content attributes, learn such a common representation in sparse user-video interactions and deal with the cold-start problem. In this paper, we propose a REgularized Dual-fActor Regression (REDAR) method based on matrix factorization. In this method, social attributes and content attributes are flexibly combined, and social and content information are effectively exploited to alleviate the sparsity problem. An incremental version of REDAR is designed to solve the cold-start problem. We extensively evaluate the proposed method for video recommendation application in real social network dataset, and the results show that, in most cases, the proposed method can achieve a relative improvement of more than 20% compared to state-of-the-art baseline methods.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

Keywords

video recommendation, attributed representation, social attributes

MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654946.

1. INTRODUCTION

Recommender systems are becoming increasingly important because of the overload of information brought about by today's Internet. Video recommender systems are required, in particular, because of the high timing costs of watching videos. Netflix reported that 75% of the content that people watch follows a recommendation. These video recommender systems therefore play the role of a bridge between users and videos, where a common representation of videos and users is required to measure the matching degree of user-video pairs. Developing an interpretable and stationary common representation method for both videos and users is of paramount significance for effective and efficient video recommendation.

In the literature, collaborative filtering (CF) has achieved great success in recommender systems. User-based CF methods represent users with videos as features, such that user-video matching can be conducted in the item space. In contrast, item-based CF methods represent videos with users as features and calculate the matching degree of user-video pairs in user space. However, the performances of these methods are seriously affected by the sparsity of the user-video collaborative matrix; they are unable to infer meaningful information about videos (or users) that lack interactions with different users (or videos). More recently, matrix factorization based CF has become more popular. It assumes a common low-dimensional latent factor representation for both users and items such that the user-item matching degree is measurable in the latent space. However, the latent factors are hardly interpretable, which makes it difficult to generalize the learned representations to new data. In addition, all CF methods suffer from the cold start problem, i.e., making recommendations for new users or new videos difficult, owing to a lack of information in the collaborative matrix. To address this, content-based video recommendation methods are currently being investigated, where users' interests and preferences are represented in detail by either content features or metadata (for example, title, or tags) of videos. However, the low-level representation is often too specific to capture users' broad interests [6]. Representing videos and users in a common space with good interpretability, stability, and appropriate levels of granularity is still an open problem.

Fortunately, the emergence of social media brought us with vast amount of users, videos and the observable interaction behaviors between users and videos. With the *Homophily* hypothesis[1], it is therefore possible and reasonable to extract a middle-level repre-

[1] Homophily is the tendency of individuals to associate and bond with others who are similar. It is often used to account for the similar behaviors of similar people towards new ideas or innovations, which implies that user-item interactions can be predicted by similarity matching of users and items in a common space.

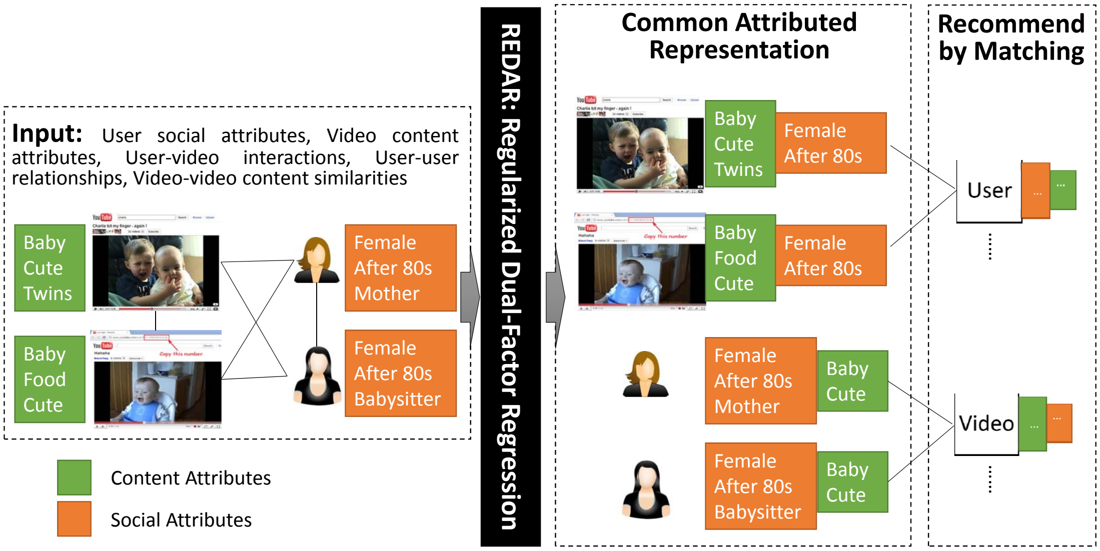
Input: User social attributes, Video content attributes, User-video interactions, User-user relationships, Video-video content similarities
Baby Cute Twins
Baby Food Cute
Female After 80s Mother
Female After 80s Babysitter
Content Attributes
Social Attributes
REDAR: Regularized Dual-Factor Regression
Common Attributed Representation
Baby Cute Twins
Female After 80s
Baby Food Cute
Female After 80s
Female After 80s Mother
Baby Cute
Female After 80s Babysitter
Baby Cute
Recommend by Matching
User
Video

Figure 1: The framework of attributed representation based video recommendation.

sentation layer from the social-enriched environment to represent both videos and users. Attributes, which lie in the middle of low-level features and high-level semantics, have been intensively investigated in recent years to characterize visual contents. Furthermore, it has been demonstrated that such a middle-level representation is an appropriate level of granularity for various application scenarios. It is natural and straightforward to extract content attributes from videos to depict their contents. In addition, we can extract social attributes from user profiles to capture users' characteristics in social aspects. Through the interactions between users and videos, the video content attributes and user social attributes are no longer isolated. If a video is watched by a number of users, the social attributes of the users who watched the video can be aggregated into the social attributes of the video. For example, in Figure 1, a video can be represented by social attributes such as "after 80s" and "female," which means that the video is more probably liked by after 80s female users. Similarly, for a user having watched a number of videos, the content attributes of the videos that are watched by the user can be aggregated into the content attributes of the user. Figure 1 shows that the user can be represented by content attributes such as "baby" and "cute," which means that the user is more likely to watch videos with such content. In this way, both social attributes and content attributes can be used as the common representation for user-video matching, and such representations are interpretable (compared with latent space models), stable (compared with CF models), and have an appropriate level of granularity.

However, it is still challenging to make video recommendations in the social and content attribute space.

(1) **The balance of social attributes and content attributes**. User-video interactions are often triggered by both long-term preference and instantaneous interest, alongside the general topic or style of the video, or detailed elements of the video content that attracted the user. Comparatively, social attribute-based representation is more stable and general. Users' social attributes are often abstract (e.g., after 80s) and remain unchanged for quite a long time. Although videos dynamically propagate among users with different social attributes, their aggregated social attributes are unlikely to dramatically vary over time. In contrast, users' content attributes can capture user preferences in detail and are therefore more dynamic with users watching videos with different content attributes. A method which integrates a user's content attributes incorporating different aspects of a user-video interaction mechanism requires a subtle design.

(2) **The sparsity of user-video interactions.** Extracting both a user's social attributes from user profiles and content attributes for videos from their surrounding texts is straightforward. However, in order to learn social attributes for videos and vice versa, user-video interactions play an important bridging role. Nevertheless, the high volumes of users and videos intrinsically determine the sparsity of user-video interactions. Therefore, effective priors are required to alleviate the sparsity problem. Embedding rich information, in addition to user-video interactions in social media, into the recommendation framework is a critical problem.

(3) **The cold-start problem.** Making recommendations for cold users (who have not watched any videos) and cold videos (that have not been watched by any users), the so-called cold-start problem, always proves difficult to overcome in video recommender systems. As mentioned above, we rely on user-video interactions to bridge users and content attributes, as well as videos and social attributes. For new users and videos, we need an effective method to link them with other active users and popular videos, whose attribute-based representations have been learned from rich information.

In this paper, we analyze the significance of social attributes for videos and content attributes for users. We find that most videos can obtain a number of representative social attributes by aggregating the social profiles of the users who watched them. This is also the case for users and content attributes. Furthermore, we find that videos having similar content are prone to have similar social attributes, and users having social relationships are prone to have similar content attributes. These interesting discoveries imply the possibility of video recommendation in social and content attribute space. Thus we further propose a REgularized Dual fActor Regression (REDAR) method based on matrix factorization. In particular, with the aim of predicting user-video interactions, we factorize the observed user-video interactions into two factors, which respectively correspond to the matching degree of user-video pairs in both social attribute space and content attribute space. These two factors are flexibly combined to optimize the approximation of the user-video interactions. In order to alleviate the sparsity problem of user-video interactions, we use video content similarities to regularize the similarities between the social attributes of videos, and we also use social relationships to regularize the similarities between the users' content attributes. For new users and videos, we propose a cold-start strategy, where new videos (or users) are efficiently linked with other videos (or users) by content similarity (or social relationship). We evaluate our proposed method on real online social media data collected from a Twitter-style website. The experiments on the real data validate our hypotheses and demonstrate the superiority of REDAR for video recommendation.

It is worthwhile to highlight the key contributions of this paper:

(1) In contrast with the representation methods in traditional recommender systems, we attempt to represent (a) videos with social attributes and (b) users with content attributes by harvesting video propagation traces among users and users' interactive behavior with videos. The resulting common attribute-based representation is interpretable with an appropriate level of granularity, in which videos can be effectively and efficiently recommended to users by directly measuring the user-video similarities (see Sections 1 and 2).

(2) We validate the rationality for representing videos (or users) with social (or content) attributes using data statistics in a real social network dataset, and find the patterns of these representations among videos with similar visual contents and users with social relationships. These discoveries pave the way for recommending videos to users in the representation space with common attributes (see Section 3).

(3) We propose our REDAR method based on matrix factorization to predict user-video interaction behavior and present video recommendations accordingly. The model can attain a superior trade-off between social attribute-based representation and content attribute-based representation, and incorporate flexible regularizers from social and visual information to alleviate the sparsity problem. In addition, a cold-start strategy for REDAR is proposed, which can also be used to efficiently deal with online and incremental data (see Section 4).

(4) We extensively evaluate the proposed methods using a reasonable scale real dataset. The experimental results show that our proposed REDAR method can significantly and consistently outperform other state-of-the-art baseline methods. Several variants of REDAR are compared and analyzed to demonstrate the rationality of the design (see Section 5).

2. RELATED WORK

In this section, we will briefly survey the related work, introduce the corresponding taxonomies, and position the uniqueness of this paper.

(1) **Traditional Recommendation Methods**

Content-based filtering and CF have been widely used to help users discover the most valuable information to them. Content-based filtering introduces the basic idea of studying the item content for the ranking problem. With the emergence of topic modeling techniques such as LDA [3], recent content-based approaches [21] rank candidate items by how well they match the topic interest of the user as their preference. These methods represent detailed users and items, enabling them to recommend similar items to what the user has previously adopted. CF methods, consisting of memory-based and model-based methods, are widely used. The memory-based approaches [20] calculate the similarity between all users based on their ratings of items. They represent users (or items) by the item-sets (or user-sets), which are often unstable and can only obtain good performance for active users or popular items. The model-based methods learn a model based on patterns recognized in the ratings of users. Several matrix factorization methods [13] have recently been proposed. The matrix approximation models all focus on representing the user-item rating matrix with low-dimensional latent vectors. These learned latent representations are hardly interpretable and are therefore difficult to generalize to new data. Although these methods have achieved success in real applications, the representation methods they adopt limit space for improvement.

(2) **Social Recommendation**

Recognizing that influence is a subtle force that governs the dynamics of social networks, influence-based recommendation [14] involves interpersonal influence into social recommendation cases. Trust-based approaches[8] exploit the trust network among users and make recommendations based on the ratings of users who are directly or indirectly trusted. [9][11] proposed a probabilistic factor analysis framework, which fuses users' preference and social influence together. Furthermore, [10] investigated the social recommendation problem in a multiple domain setting. Most of these works are based on traditional content-based filtering or CF-based filtering methods, and their common goal is to embed social information into traditional methods to improve the recommendation accuracy. However, few authors have targeted the problem of how to learn a new common representation for users and items in social networks, which is indeed feasible and important for boosting social recommendation performance.

(3) **Video Recommendation**

Video recommendation plays an important role in delivering videos to users. The most well-known real video recommendation system is running on YouTube, where personalized sets of videos are recommended to users based on their activities on the site [5]. The recommendation system is one of the most important techniques to find videos [29], preceded by video searching. Compared with YouTube's recommendation system, which only considers user ratings and user-video interaction information, the following research on video recommendation attempted to incorporate more aspects of user and content information[23]. Park et al. [19] proposed to construct user profiles as an aggregation of tag clouds and generate recommendations according to similar viewing patterns. Bertini et al. [2] built a demo to show how to create user profiles containing the users' interests and apply them to a friend's suggestion and video recommendation. Ma et al. [16] assumed that social friends have higher common interests and their sharing behaviors are an important clue to enhance video recommendation. These methods pay more attention to user representation by discovering user profiles and behavior patterns [24]. Further, Zhu et al. [30] decomposed the recommendation process into video representation and recommendation generation, and represented videos with topics to match user interests. Fu et al. [7] tackled the problem of attribute learning for videos with sparse labels. These methods focus on video representation, yet the representations of videos and users are still isolated [18, 26]. Previous works focused on either user representation or video representation, and only a few works investigated integration strategies [25, 15] by first separately processing and then aggregating for recommendation. Our work is the first to uniformly learn a common representation for both users and videos in an interpretable way and with proper granularity.

(4) **Attribute based Methods**

Recently, attributes have aroused much interest in the research community because attribute-based representations are in a proper granularity to bridge low-level information and high-level semantics. Khalid et al. [6] represented documents by attributes of users

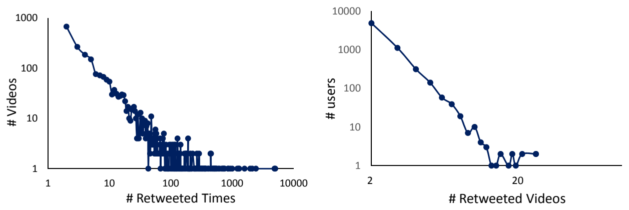

Figure 2: Distribution of users (or videos) with respect to the number of videos they retweeted (or the number of times they are retweeted).

who read these documents. Zhang et al. [28] used an attribute-augmented semantic hierarchy to bridge the semantic gap and intention gap. Ma et al. [17] extracted visual, audio, and linguistic attributes for video summarization, which outperforms traditional visual-only methods, while Cha et al. [4] conducted statistical analysis on video attributes to analyze user behavior on Youtube, giving an intuitive observation of the video sharing platform. Yu et al. [27] regarded user behaviors as their attributes to explore large-scale video-on-demand systems. These methods investigated attribute representations either for videos or for users. However, only a limited number of them explored the possibility of a common attribute-based representation for both users and videos, which is the focus of this paper.

3. PRELIMINARY STUDY

In this section, we will introduce the characteristics of the dataset and validate the rationality of representing videos (or users) by social attributes (or content attributes).

3.1 Data Description

The dataset is collected from Tencent Weibo, a Twitter-style social network platform in China with more than 300 million users. Video sharing is an important and popular feature of this platform. Users proactively forward the videos from external Youtube-style video sharing websites onto this microblog platform; the video then propagates over the social graph. We collected these kinds of videos with their text descriptions between June 20, 2012 and June 26, 2012. We then selected the videos that had been retweeted at least two times, collected the profiles (including demographic information and self-labeled tags such as "after 90s," "soccer fans," etc.) of the users that were involved in the propagation processes of these videos at least two times, and the social relations among the collected users. Therefore, we have 2357 videos, 6572 users, and 1271 social relations among the users. We show the distribution of videos with respect to the times they are retweeted, and the distribution of users with respect to the number of videos that they retweeted in Figure 2. It can be seen that both distributions obey the power law, which indicates that the user-video interactions are sparse. Thus, the dataset is adequate for simulating real application scenarios and evaluating the performances of the proposed method and other baselines.

3.2 Social Attributes of Videos

Here, we use a simple method to intuitively approximate the social attribute-based representations for videos, while a formal method will be proposed in Section 4 under the motivation of this preliminary study. We first extract social attributes for users. The demographic information is easily transformed into labels. For the self-labeled tags, we filter the tags with very low or very high frequency, while retaining the mid-frequency tags as labels. All these labels are used as the social attribute candidate set of users. For each video, we aggregate the social attributes of the users who retweeted the video. If the frequency of a social attribute exceeds 20% of the total number of users who retweeted the video, this attribute is regarded as a valid social attribute for the video. The distribution of videos with respect to the number of valid social attributes is shown in Figure 3. It can be seen that the agglomeration phenomenon of social attributes universally exists among videos, which validates the hypothesis of videos being watched by users with similar social attributes. However, the figure also shows that most videos only have a small number of valid social attributes, which pose a significant challenge to video recommendation based on such a sparse representation.

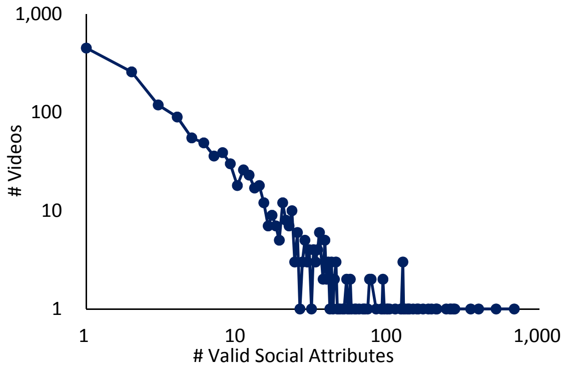

Figure 3: Distribution of videos with respect to the number of valid social attributes.

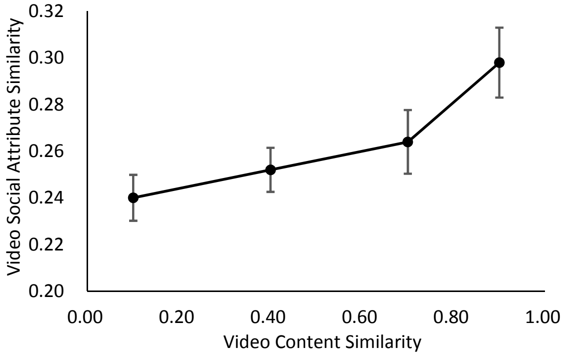

Figure 4: The correlation between videos' social attribute similarity and their content similarity.

There is a well-accepted belief that a user is more likely to like videos that are similar to those videos they have watched previously. This is the fundamental hypothesis of content-based video recommendation methods. Motivated by this, we presume that videos with similar visual contents should have similar valid social attributes. To validate this, we calculate the visual content similarity of each pair of videos (the detail method for calculating the visual content similarity is referred to in Section 4.1) and the social attribute similarity of each pair of videos, and plot the relationship between the visual content similarity and the social attribute similarity, as shown in Figure 4. It can be seen that, on average, social attribute similarities of videos positively correlates with their visual content similarities. That is, videos that have similar visual contents are more likely to have similar social attributes.

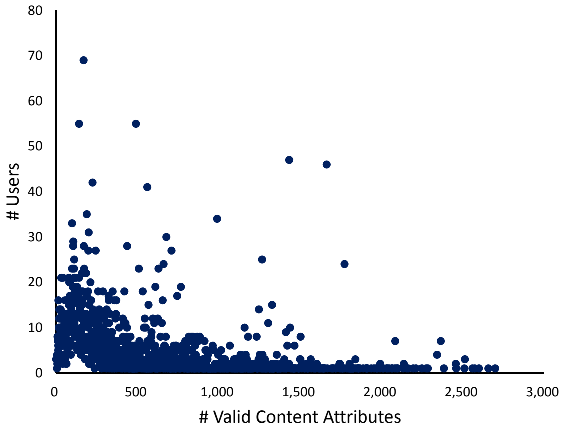

Figure 5: Distribution of users with respect to the number of valid content attributes.

3.3 Content Attributes of Users

We deal with the content attributes of users in a similar approach. We first extract content attributes for videos. Almost all videos have text descriptions to indicate the high-level semantics of the video contents. These text descriptions are segmented into words, and after filtering the words with very low or very high frequency, we retain the mid-frequency words as labels. All these labels are used as the content attribute candidate set of videos. For each user, we aggregate the content attributes of the videos that the user watched. If the frequency of a content attribute exceeds 20% of the total number of videos that the user watched, this attribute is regarded as a valid content attribute for the user. The distribution of videos with respect to the number of valid social attributes is shown in Figure 5. Similarly, the agglomeration phenomenon of content attributes universally exists among users, yet the valid content attributes for each user are very sparse.

The homophily hypothesis is a commonly accepted hypothesis in sociology to interpret the interpersonal relations within social networks. It presumes that users with similar demographic information or attributes are more prone to establish social relations. In our case, we use the hypothesis conversely and presume that users with social relations are more likely to have similar content attributes. To validate this, we measure the content attribute similarities of all pairs of users, and calculate the average and standard deviation of content attribute similarity for the pairs of users with and without social relations, respectively. We show the results in Figure 6, and it can be seen that, on average, the content attribute similarities of users who are socially linked are much higher than those of users without social relations.

We have now validated the important hypotheses in this work, including (1) the agglomeration phenomenon exists in social attributes among videos and content attributes among users; (2) videos with similar visual contents are more likely to have similar social attributes; and (3) users with social relations are more likely to have similar content attributes. The first one is fundamental for representing videos (or users) with social attributes (or content attributes), while the latter two provide important clues for alleviating the sparsity and cold-start problem in attribute-based representation video recommendation.

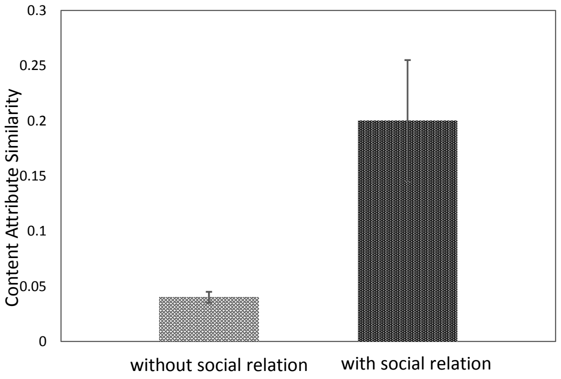

Figure 6: The correlation between users' content attribute similarity and their social relationship.

4. THE REDAR MODEL

4.1 Problem Formulation

Suppose that we have collected abundant information from a social media platform, with large sets of users $\mathcal{U}$ and videos $\mathcal{V}$ with $M = |\mathcal{U}|$ and $N = |\mathcal{V}|$. We denote the watching relationship matrix as $\mathbf{W} \in \{0,1\}^{M\times N}$, with its (u,v)-th entry

$$W_{uv} = \begin{cases} 1, & \text{if user } u \text{ watched video } v; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

However, in real cases, most of the elements in $\mathbf{W}$ are zero because of the sparse interactions between users and videos. Thus, in order to focus more on the valid elements, we propose to only measure the approximation loss on observed elements on $\mathbf{W}$. To formulate this, we introduce the *masking matrix* $\mathbf{Y} \in \{0,1\}^{M\times N}$, with its (u,v)-th entry

$$Y_{uv} = \begin{cases} 1, & \text{if user } u\text{'s behavior on video } v \text{ is observed;} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where observed behaviors could be either positive or negative. Please refer to Section 5.1 for details on how to determine whether an entry is observed.

We denote the the friendship relationship matrix as $\mathbf{R} \in \{0,1\}^{M\times M}$, with its (u_1,u_2)-th entry

$$R_{u_1u_2} = \begin{cases} 1, & \text{if user } u_1 \text{ is a friend of user } u_2; \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We denote the content similarity matrix as $\mathbf{C} \in [0,1]^{N\times N}$, with its (v_1,v_2)-th entry

$$C_{v_1v_2} = e^{-distance(v_1,v_2)} \tag{4}$$

where $distance(v_1,v_2)$ is measured by visual features. Given two videos, keyframes are extracted respectively, with SIFT descriptors pre-extracted on these keyframes, so that it won't consume a lot of online calculations. Keyframes from the two videos are matched to count the matched number of descriptors, and the total ratio of matched descriptors is used as the video distance [12].

Besides visual features, tags are also considered in social media to describe videos and users. Tags of a video come from accompanying text of the video. The video tag set is denoted as $\mathcal{T}^{\mathcal{V}}$ with $P = |\mathcal{T}^{\mathcal{V}}|$, and the relationships of videos and video tags are denoted by a matrix $\mathbf{V} \in [0,1]^{N\times P}$, with

$$V_{v,t} = TF_v(v,t) \cdot IDF_v(t) \tag{5}$$

where TF_v is calculated on video v's accompanying text, and IDF_v is calculated on all video accompanying texts. Tags of a user can be extracted from the user's profile and his/her self-labeling tags. The user tag set is denoted as $\mathcal{T}^{\mathcal{U}}$ with $Q = |\mathcal{T}^{\mathcal{U}}|$, and the relationship of users and user tags is denoted by a matrix $\mathbf{U} \in [0,1]^{M \times Q}$, with

$$U_{u,t} = TF_u(u,t) \cdot IDF_u(t) \tag{6}$$

where TF_u is calculated on user u's profile, and IDF_u is calculated on all user profiles.

With all these matrices as input, according to the observations in Section 3, we can define our targeting matrices. The first targeting matrix is $\mathbf{E} \in [0,1]^{M \times P}$, which links users with videos' attributes. The second targeting matrix is $\mathbf{F} \in [0,1]^{N \times Q}$, which links videos with users' social attributes. In order to learn the targeting matrices from the input information, we use the following principles to define our objective function:

- User-video matching with both social attributes and content attributes should be consistent with the real user-video interactions.
- User-user similarity with content attributes should be consistent with user friendship.
- Video-video similarity with social attributes should be consistent with videos' visual similarity.
- Both social attributes and content attributes should be sparse.

Therefore, the objective function is defined as

$$\begin{aligned}(\mathbf{E},\mathbf{F}) = \arg\min_{\mathbf{E},\mathbf{F}} (&||\mathbf{Y} \odot (\alpha \mathbf{E}\mathbf{V}^T + (1-\alpha)\mathbf{U}\mathbf{F}^T - \mathbf{W})||_F^2 \\ &+ \lambda_1 ||\mathbf{E}\mathbf{E}^T - \mathbf{R}||_F^2 + \lambda_2 ||\mathbf{F}\mathbf{F}^T - \mathbf{C}||_F^2 \\ &+ \lambda_3 ||\mathbf{E}||_1 + \lambda_4 ||\mathbf{F}||_1)\end{aligned} \tag{7}$$

where $\alpha \in [0,1]$ is a weighting factor between video tags and user tags, and λ_1, λ_2, λ_3, λ_4 are non-negative parameters less than 1, with λ_1 weighting for user relationships and λ_2 weighting for visual similarities. $\odot$ stands for an element by element multiply operation. Although there are five parameters to tune in this model, $alpha$, λ_1 and λ_2 are more important, and the result is not sensitive with λ_3 and λ_4. In this objective function, both social attribute based representation and content attribute based representation are flexibly integrated, and rich side information are embedded into regularizers to alleviate the sparsity problem. Thus we denote the proposed method as *REgularized Dual-fActor Regression* (REDAR).

4.2 The Model Solution

To solve the problem, we apply Iterative Shrinkage Thresholding Algorithm [1]. To simplify the description, we define

$$\begin{aligned}f(\mathbf{E},\mathbf{F}) =& ||\mathbf{Y} \odot (\alpha \mathbf{E}\mathbf{V}^T + (1-\alpha)\mathbf{U}\mathbf{F}^T - \mathbf{W})||_F^2 \\ &+ \lambda_1 ||\mathbf{E}\mathbf{E}^T - \mathbf{R}||_F^2 + \lambda_2 ||\mathbf{F}\mathbf{F}^T - \mathbf{C}||_F^2\end{aligned} \tag{8}$$

$$g(\mathbf{E},\mathbf{F}) = \lambda_3 ||\mathbf{E}||_1 + \lambda_4 ||\mathbf{F}||_1 \tag{9}$$

So that the iterative formulas to solve the model are

$$\mathbf{E}^{(k)} = T_{\lambda_3 t_k^E}(\mathbf{E}^{(k-1)} - t_k^E \nabla_{\mathbf{E}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)})) \tag{10}$$

$$\mathbf{F}^{(k)} = T_{\lambda_4 t_k^F}(\mathbf{F}^{(k-1)} - t_k^F \nabla_{\mathbf{F}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)})) \tag{11}$$

where $\mathbf{E}^{(k)}$, $\mathbf{F}^{(k)}$ are the k-th round of results of our targeting matrices $\mathbf{E}$ and $\mathbf{F}$. t_k^F and t_k^E are step sizes. The function T_λ is defined as

$$T_\lambda(x) = (|x| - \lambda)_+ sgn(x) \tag{12}$$

In order to solve $\nabla_{\mathbf{E}} f$ and $\nabla_{\mathbf{F}} f$, we have

$$f(\mathbf{E},\mathbf{F}) = f_1(\mathbf{E},\mathbf{F}) + \lambda_1 f_2(\mathbf{E}) + \lambda_2 f_3(\mathbf{F}) \tag{13}$$

where

$$f_1(\mathbf{E},\mathbf{F}) = ||\mathbf{Y} \odot (\alpha \mathbf{E}\mathbf{V}^T + (1-\alpha)\mathbf{U}\mathbf{F}^T - \mathbf{W})||_F^2 \tag{14}$$

$$f_2(\mathbf{E}) = ||\mathbf{E}\mathbf{E}^T - \mathbf{R}||_F^2 \tag{15}$$

and

$$f_3(\mathbf{F}) = ||\mathbf{F}\mathbf{F}^T - \mathbf{C}||_F^2 \tag{16}$$

By solving their derivations, we get

$$\frac{\partial f_1}{\partial \mathbf{E}} = 2\alpha(\mathbf{Y} \odot (\alpha \mathbf{E}\mathbf{V}^T + (1-\alpha)\mathbf{U}\mathbf{F}^T - \mathbf{W}))\mathbf{V} \tag{17}$$

$$\frac{\partial f_1}{\partial \mathbf{F}} = 2(1-\alpha)(\mathbf{Y}^T \odot (\alpha \mathbf{E}\mathbf{V}^T + (1-\alpha)\mathbf{U}\mathbf{F}^T - \mathbf{W})^T)\mathbf{U} \tag{18}$$

$$\frac{\partial f_2}{\partial \mathbf{E}} = 4(\mathbf{E}\mathbf{E}^T - \mathbf{R})\mathbf{E} \tag{19}$$

$$\frac{\partial f_3}{\partial \mathbf{F}} = 4(\mathbf{F}\mathbf{F}^T - \mathbf{C})\mathbf{F} \tag{20}$$

Therefore, we get

$$\nabla_{\mathbf{E}} f(\mathbf{E},\mathbf{F}) = \frac{\partial f_1}{\partial \mathbf{E}} + \lambda_1 \frac{\partial f_2}{\partial \mathbf{E}} \tag{21}$$

$$\nabla_{\mathbf{F}} f(\mathbf{E},\mathbf{F}) = \frac{\partial f_1}{\partial \mathbf{F}} + \lambda_2 \frac{\partial f_3}{\partial \mathbf{F}} \tag{22}$$

More details about the flow of the algorithm is presented in Algorithm 1.

Algorithm 1 Iterative Solution of REDAR Model.

```
Input: $\mathbf{W}, \mathbf{R}, \mathbf{C}, \mathbf{U}, \mathbf{V}$
       $t_k^E, t_k^F, \lambda_1, \lambda_2, \lambda_3, \lambda_4$
Output: $\mathbf{E}, \mathbf{F}$
Initialize: $\mathbf{E}^{(0)} = \mathbf{W}\mathbf{V}, \mathbf{F}^{(0)} = \mathbf{W}^T\mathbf{U}$
for $k = 1, 2, \ldots$ do
    Calculate $\frac{\partial f_1^{(k-1)}}{\partial \mathbf{E}}, \frac{\partial f_1^{(k-1)}}{\partial \mathbf{F}}, \frac{\partial f_2^{(k-1)}}{\partial \mathbf{E}}, \frac{\partial f_3^{(k-1)}}{\partial \mathbf{F}}$
    $\nabla_{\mathbf{E}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)}) = \frac{\partial f_1^{(k-1)}}{\partial \mathbf{E}} + \lambda_1 \frac{\partial f_2^{(k-1)}}{\partial \mathbf{E}}$
    $\nabla_{\mathbf{F}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)}) = \frac{\partial f_1^{(k-1)}}{\partial \mathbf{F}} + \lambda_2 \frac{\partial f_3^{(k-1)}}{\partial \mathbf{F}}$
    $\mathbf{E}^{(k)} = T_{\lambda_3 t_k^E}(\mathbf{E}^{(k-1)} - t_k^E \nabla_{\mathbf{E}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)}))$
    $\mathbf{F}^{(k)} = T_{\lambda_4 t_k^F}(\mathbf{F}^{(k-1)} - t_k^F \nabla_{\mathbf{F}} f(\mathbf{E}^{(k-1)}, \mathbf{F}^{(k-1)}))$
end for
return $\mathbf{E}^{(k)}, \mathbf{F}^{(k)}$
```

With the optimized results of $\mathbf{E}$ and $\mathbf{F}$, which are denoted as $\hat{\mathbf{E}}$ and $\hat{\mathbf{F}}$, respectively, we can estimate the user-video interactions by

$$\hat{\mathbf{W}} = \alpha \hat{\mathbf{E}}\mathbf{V}^T + (1-\alpha)\mathbf{U}\hat{\mathbf{F}}^T \tag{23}$$

Time complexity analysis. The time complexity of the algorithm depends on the matrix calculations and number of iterations. Since both matrix linear combinations and lasso take linear time, we only need to consider matrix calculation of the gradients. More specifically, the running time is

$$\begin{aligned}&T(k,M,N,P,Q) \\ =&O(k)(O(\frac{\partial f_1}{\partial \mathbf{E}}) + O(\frac{\partial f_1}{\partial \mathbf{F}}) + O(\frac{\partial f_2}{\partial \mathbf{E}}) + O(\frac{\partial f_3}{\partial \mathbf{F}})) \\ =&O(k(M(P+Q)N + MNP + MNQ + MPM + NQN)) \\ =&O(k((M+N)(M+N)(P+Q) - MMQ - NNP)) \\ \leq&O(k(M+N)(||\mathbf{E}||_0 + ||\mathbf{F}||_0 + ||\mathbf{U}||_0 + ||\mathbf{V}||_0))\end{aligned} \tag{24}$$

We use MP, NQ, MQ, and NP to denote the 0-norm of $\mathbf{E}$, $\mathbf{F}$, $\mathbf{U}$, and $\mathbf{V}$, respectively, which stands for the number of non-zero elements in the matrix. That is to say, our algorithm can compute $||\mathbf{E}||_0 + ||\mathbf{F}||_0$ and scan $||\mathbf{U}||_0 + ||\mathbf{V}||_0$ elements with each in time of $O(k(M+N))$ on average. Further, with the scope matrix $\mathbf{Y}$, the matrix calculation can avoid the calculation of most elements, which results in a small const factor of the time complexity.

4.3 Incremental REDAR Model

Our model can also handle the cold-start problem with an incremental algorithm. To deal with new users and new videos, we do not learn the entire $\mathbf{E}$ and $\mathbf{F}$, but only focus on the incremental part of the data. More specifically, suppose we have a set of new users $\Delta\mathcal{U}$ with $\Delta M = |\Delta\mathcal{U}|$ and a set of new videos $\Delta\mathcal{V}$ with $\Delta N = |\Delta\mathcal{V}|$. We do not know their watching histories, i.e., the relationships between the new users and the new videos. We represent the friendship of the new users and all users by a matrix $\Delta\mathbf{R} \in \{1,0\}^{(M+\Delta M)\times\Delta M}$, with the same entity meaning to $\mathbf{R}$. We represent the visual feature similarity of new videos and all videos as a matrix $\Delta\mathbf{C} \in [0,1]^{(N+\Delta N)\times\Delta N}$, with the same entity meaning to $\mathbf{C}$. We denote new video tag set by $\Delta\mathbf{V} \in [0,1]^{\Delta N\times P}$, with the same entity meaning to $\mathbf{V}$. We denote new user tag set by $\Delta\mathbf{U} \in [0,1]^{\Delta M\times Q}$, with the same entity meaning to $\mathbf{U}$. Also, we denote our targeting matrices by $\Delta\mathbf{E} \in [0,1]^{\Delta M\times P}$ and $\Delta\mathbf{F} \in [0,1]^{\Delta N\times Q}$, with the same entity meaning to $\mathbf{E}$ and $\mathbf{F}$, respectively.

Since the missing of new user-video watching relationship, we only need to focus on the friendship and visual feature similarity terms. Thus we get the following objective function

$$\begin{aligned}(\Delta\mathbf{E},\Delta\mathbf{F}) = \arg\min_{\Delta\mathbf{E},\Delta\mathbf{F}} \lambda_1 || \begin{bmatrix} \mathbf{E} \\ \Delta\mathbf{E} \end{bmatrix} \times \Delta\mathbf{E}^T - \Delta\mathbf{R}||_F^2 \quad (25)\\ +\lambda_2 || \begin{bmatrix} \mathbf{F} \\ \Delta\mathbf{F} \end{bmatrix} \times \Delta\mathbf{F}^T - \Delta\mathbf{C}||_F^2 \\ +\lambda_3||\Delta\mathbf{E}||_1 + \lambda_4||\Delta\mathbf{F}||_1\end{aligned}$$

which can be solved in a similar way with Equation 7 as follows. We define

$$\begin{aligned}\Delta f(\Delta\mathbf{E},\Delta\mathbf{F}) = \lambda_1 || \begin{bmatrix} \mathbf{E} \\ \Delta\mathbf{E} \end{bmatrix} \times \Delta\mathbf{E}^T - \Delta\mathbf{R}||_F^2 \quad (26)\\ + \lambda_2 || \begin{bmatrix} \mathbf{F} \\ \Delta\mathbf{F} \end{bmatrix} \times \Delta\mathbf{F}^T - \Delta\mathbf{C}||_F^2\end{aligned}$$

$$\Delta g(\Delta\mathbf{E},\Delta\mathbf{F}) = \lambda_3||\Delta\mathbf{E}||_1 + \lambda_4||\Delta\mathbf{F}||_1 \quad (27)$$

So that the iterative solution is:

$$\Delta\mathbf{E}^{(k)} = T_{\lambda_3 t_k^{\Delta E}}(\Delta\mathbf{E}^{(k-1)} - t_k^{\Delta E}\nabla_{\Delta\mathbf{E}} f(\Delta\mathbf{E}^{(k-1)},\Delta\mathbf{F}^{(k-1)})) \quad (28)$$

$$\Delta\mathbf{F}^{(k)} = T_{\lambda_4 t_k^{\Delta F}}(\Delta\mathbf{F}^{(k-1)} - t_k^{\Delta F}\nabla_{\Delta\mathbf{F}} f(\Delta\mathbf{E}^{(k-1)},\Delta\mathbf{F}^{(k-1)})) \quad (29)$$

To solve $\nabla_{\Delta\mathbf{E}} f$ and $\nabla_{\Delta\mathbf{F}} f$, we have

$$\Delta f(\Delta\mathbf{E},\Delta\mathbf{F}) = \lambda_1 \Delta f_1(\Delta\mathbf{E}) + \lambda_2 \Delta f_2(\Delta\mathbf{F}) \quad (30)$$

where

$$\Delta f_1(\Delta\mathbf{E}) = || \begin{bmatrix} \mathbf{E} \\ \Delta\mathbf{E} \end{bmatrix} \times \Delta\mathbf{E}^T - \Delta\mathbf{R}||_F^2 \quad (31)$$

$$\Delta f_2(\Delta\mathbf{F}) = || \begin{bmatrix} \mathbf{F} \\ \Delta\mathbf{F} \end{bmatrix} \times \Delta\mathbf{F}^T - \Delta\mathbf{C}||_F^2 \quad (32)$$

By solving their derivations, we get

$$\frac{\partial\Delta f_1}{\partial\Delta\mathbf{E}} = 2(\begin{bmatrix} \mathbf{E} \\ \Delta\mathbf{E} \end{bmatrix} \times \Delta\mathbf{E}^T - \Delta\mathbf{R})^T \begin{bmatrix} \mathbf{E} \\ 2\Delta\mathbf{E} \end{bmatrix} \quad (33)$$

$$\frac{\partial\Delta f_2}{\partial\Delta\mathbf{F}} = 2(\begin{bmatrix} \mathbf{F} \\ \Delta\mathbf{F} \end{bmatrix} \times \Delta\mathbf{F}^T - \Delta\mathbf{C})^T \begin{bmatrix} \mathbf{F} \\ 2\Delta\mathbf{F} \end{bmatrix} \quad (34)$$

Therefore, we get

$$\nabla_{\Delta\mathbf{E}} f = \lambda_1 \frac{\partial\Delta f_1}{\partial\Delta\mathbf{E}} \quad (35)$$

$$\nabla_{\Delta\mathbf{F}} f = \lambda_2 \frac{\partial\Delta f_2}{\partial\Delta\mathbf{F}} \quad (36)$$

With the optimized $\Delta\mathbf{E}$ and $\Delta\mathbf{F}$, which are denoted by $\Delta\hat{\mathbf{E}}$ and $\Delta\hat{\mathbf{F}}$, respectively, we can estimate the user-video watching relationship as $\Delta\hat{\mathbf{W}}^{\Delta\mathcal{U}} \in [0,1]^{\Delta M\times(N+\Delta N)}$ and $\Delta\hat{\mathbf{W}}^{\Delta\mathcal{V}} \in [0,1]^{(M+\Delta M)\times\Delta N}$ for new users and new videos, respectively, where

$$\Delta\hat{\mathbf{W}}^{\Delta\mathcal{U}} = \alpha\Delta\hat{\mathbf{E}} \begin{bmatrix} \mathbf{V} \\ \Delta\mathbf{V} \end{bmatrix}^T + (1-\alpha)\Delta\mathbf{U} \begin{bmatrix} \hat{\mathbf{F}} \\ \Delta\hat{\mathbf{F}} \end{bmatrix}^T \quad (37)$$

$$\Delta\hat{\mathbf{W}}^{\Delta\mathcal{V}} = \alpha \begin{bmatrix} \hat{\mathbf{E}} \\ \Delta\hat{\mathbf{E}} \end{bmatrix} \Delta\mathbf{V}^T + (1-\alpha) \begin{bmatrix} \mathbf{U} \\ \Delta\mathbf{U} \end{bmatrix} \Delta\hat{\mathbf{F}}^T \quad (38)$$

Note that $\Delta\hat{\mathbf{W}}^{\Delta\mathcal{U}}$ and $\Delta\hat{\mathbf{W}}^{\Delta\mathcal{V}}$ are consistent in their common elements.

Time complexity analysis. With the entire data, we have to process matrices with sizes MN, MQ, MP, NQ, and NP. With the incremental data, we only need to process matrices with sizes $\Delta M \times Q$, $\Delta M \times P$, $\Delta N \times Q$, and $\Delta N \times P$. When $\Delta M \ll M$ and $\Delta N \ll N$, the incremental algorithm is much faster than a re-calculation approach. More detailedly, the time complexity of the incremental algorithm can be solved in a similar way with Algorithm 1.

$$\begin{aligned}&T(k,M,N,\Delta M,\Delta N,P,Q) \quad (39)\\ =&O(k)(O(\frac{\partial\Delta f_1}{\partial\Delta\mathbf{E}}) + O(\frac{\partial\Delta f_2}{\partial\Delta\mathbf{F}})) = O(k(MP\Delta M + NQ\Delta N))\\ =&O(k(M\times||\Delta\mathbf{E}||_0 + N\times||\Delta\mathbf{F}||_0))\end{aligned}$$

which keeps the low average computation time for each output element.

5. EXPERIMENTS

In this section, we will present the empirical study results on the REDAR method for video recommendation.

5.1 Experimental Settings

Training and Testing. For all recommendation methods in a social network environment, there is an inevitable problem of unobservable negative samples. In $\mathbf{W}$, $\mathbf{W}_{ij} = 1$ is used to indicate that the user i was interested in video j. However, $\mathbf{W}_{ij} = 0$ does not necessarily indicate that user i was not interested in video j, but rather user i never received and saw video j. To guarantee that all the 0-entries are correct signals that reflect the users decision to reject retweeting the videos after seeing them, we estimate the online sessions of users according to their behaviors reflected by the 1-entries of $\mathbf{W}$ as in [9]. We suppose that the users should be able to see all the videos that they received during online sessions. For example, if a user retweets a video at time t, then we suppose that the user can see all the videos they received from their friends during $[t-\Delta t, t+\Delta t]$. In our case, Δt is set to five minutes. Thus,

$\mathbf{W}_{ij} = 0$ is valid only if video j is received by user i during their online session, which is controlled by $\mathbf{Y}$. We randomly select 80% of the observed entries in $\mathbf{W}$ as training data, and use the remaining entries as testing data. The random selection is conducted 10 times, and the average results are reported.

Groundtruth. According to the above training and testing strategy, the testing entries are in actual known. Therefore, we use the actual value of these entries as the ground truth to evaluate the testing performance.

Evaluation Criteria.In the following experiments, we use the RMSE (Root Mean Square Error) and MAE (Mean Average Error) to calculate the reconstruction loss of $\mathbf{W}$ on testing entries, which evaluates the prediction performance in the value aspect. We also evaluate the ranking performance using Kendall's and Spearman's ranking coefficients. Finally, we calculate the Precision@K to evaluate the prediction accuracy of the top recommended videos, which is important in real recommendation applications.

5.2 Baselines

In order to demonstrate the advantages and characteristics of the proposed method, we implemented the following four state-of-the-art methods and five variants of REDAR.

- *UserCF*. Implemented according to [22], where only user-video interaction information is used.
- *ItemCF*. Implemented according to [20], where only user-video interaction information is used.
- *SVD-based CF*. Implemented according to [13], where only user-video interaction information is used.
- *User-Label-Item*. Implemented according to [6]. A video is represented by aggregating the labels of the users who watched the video, and the video recommendation is conducted by matching video label representation and user labels.
- *REDAR-SocialAttribute*[2]. Implemented by setting $\alpha = 1$ in Equation 7.
- *REDAR-ContentAttribute*. Implemented by setting $\alpha = 0$ in Equation 7.
- *REDAR-SocialRegularizer*. Implemented by setting $\lambda_1 = 0$ in Equation 7.
- *REDAR-ContentRegularizer*. Implemented by setting $\lambda_2 = 0$ in Equation 7.
- *REDAR-LassoRegularizer*. Implemented by setting $\lambda_3 = \lambda_4 = 0$ in Equation 7.

5.3 Parameters Setting

We have five parameters in REDAR in total. α is the trade-off parameter for balancing social attribute-based representation and content attribute-based representation. $\lambda_1 - \lambda_4$ respectively control the weight of the social relation prior, the visual content prior, and the sparse prior on the social attributes of the videos and the content attributes of the users. For the parameter setting, we use grid search to obtain the optimal parameters. To demonstrate the importance of both social attributes and content attributes, we show how α affects the performance of REDAR in Figure 7. It can be seen that the RMSE of testing data significantly varies with different α values, which indicates the importance of subtly balancing social attributes and content attributes. The proposed method attains the best performance when $\alpha = 0.3$, which also indicates that, comparatively, the social attribute-based representation is more effective for video recommendation, while the role of content attributes cannot be ignored.

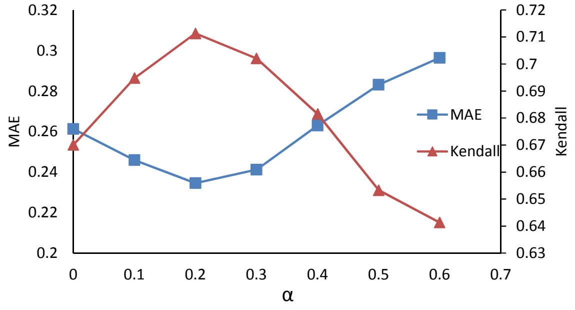

Figure 7: The parameter tuning of α to balance social attributes and content attributes.

5.4 Video Recommendation Performances

Firstly, we evaluate the prediction accuracy of the proposed method with the measures including prediction errors (MAE and RMSE) and ranking coefficients (Kendall and Spearman). As shown in Table 1, the proposed REDAR method achieves the best performance across all measures. In addition, the relative improvement of REDAR from other baseline methods is significant. For example, in RMSE, REDAR achieves 25% relative improvement compared to the best baseline method *User-based Representation*. By comparing the results of baseline methods and REDAR, we arrive at the following observations.

	MAE	RMSE	Kendall	Spearman
UserCF	0.3664	0.5744	0.5430	0.5961
ItemCF	0.3958	0.6067	0.5082	0.5578
SVD-CF	0.3204	0.5412	0.5858	0.6360
User-Label-Item	0.3104	0.5366	0.6189	0.6910
REDAR	**0.2345**	**0.4299**	**0.7113**	**0.8539**

Table 1: Video recommendation performances.

(1) *UserCF*, *ItemCF* and *SVD-based CF* depend only on the collaborative matrix information, yet *SVD-based CF* performs much better than *UserCF* and *ItemCF*. The significant improvement should be attributed to the learned latent representation for users and videos, which is more optimal than item-based or user-based representation in approximating user-video interactions.

(2) The *User-Label-Item* method performs better than *SVD-based CF*. The main reason is that the common representation of items and users learned in the *User-Label-Item* method is interpretable and stationary. Another possible explanation is that User-Label-Item method incorporates richer information than *SVD-based CF*, such as the user profiles.

(3) Although the ultimate goal of both REDAR and *User-Label-Item* is to learn a common representation of videos and users with good interpretability, stability, and appropriate granularity, the results show that REDAR performs better than *User-Label-Item*. After translating *User-Label-Item* into our scenario, *User-Label-Item* only considers the social attributes of the videos, but ignores the importance of the users' content attributes. This makes the learned

[2] In all variants of REDAR, "-" means "minus".

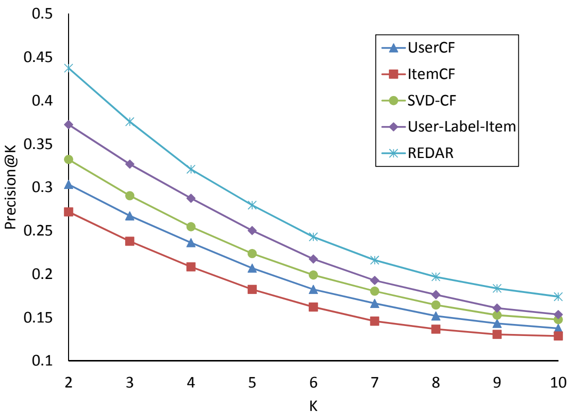

Figure 8: Performance evaluation in Precision@K.

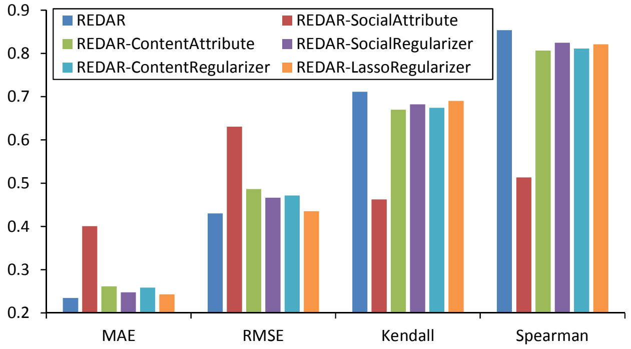

Figure 9: Performance evaluation on the variants of REDAR method.

representation difficult to capture the dynamic characteristics of users' preferences. Moreover, the REDAR method is more flexible at incorporating effective priors, such as the social relationship information.

We also evaluate the methods in Precision@K, and show the results in Figure 8. We observe that REDAR performs the best in all testing cases, and the improvement is more obvious when K is reduced. This is important in real recommendation applications, where users often only browse the top recommended videos.

5.5 Effects of Regularizers

In order to understand the characteristics and demonstrate the rationality of REDAR, we evaluate the performances of several variants of the REDAR model with a leave-one-out strategy. That is, we remove the factors and regularizers one by one and observe the change of model performances. The degree of performance degradation indicates the importance and effectiveness of the removed factor or regularizer. The results are shown in Figure 9. From the results, we can reach several observations.

(1) The largest degree of performance degradation occurs when changing REDAR into REDAR-SocialAttribute by removing the social attribute factor, which means that representing videos with social attributes is effective in video recommendation scenarios. In comparison to visual feature-based video representation and surrounding text-based video representation methods, social attribute-based video representation is at an appropriate granularity. In addition, it can avoid the semantic gap problem of low-level features and the problems brought by the absence of surrounding texts. More importantly, social attribute-based video representation can directly account for users' interaction behaviors on videos.

(2) Although content attribute-based representation produces the worst results, it complements the social attribute-based representation method, meaning they can achieve better results together than they can individually.

(3) All the regularizers imposed to the main regression term play important roles in addressing the sparsity problem. Comparatively, the social regularizer is more important than the content regularizer, which is reasonable because the social influence factor embedded in social relations is important for predicting user behaviors in a social network environment [9]. Nevertheless, the non-trivial improvement from REDAR-ContentRegularizer to REDAR justifies the contribution of visual content, which is also demonstrated in our preliminary study.

5.6 Cold-Start Recommendation Performances

In this subsection, we analyze the capability of REDAR to deal with new users and videos. We first randomly select 10% of the users in our dataset as new users, and then hide the historical video retweeting behaviors from **W** and use the data of the remaining users to learn **E** and **F**. After that, we apply our incremental version ΔREDAR to learn representations for the new users, recommend videos to these new users accordingly, and evaluate the performances of these recommendations. To demonstrate the performance of ΔREDAR, we also evaluate the performance of REDAR on these new users. We deal with new videos in a similar way to new users.

	MAE	RMSE	Kendall	Spearman	Time
REDAR	0.2362	0.4338	0.7094	0.8522	4.16h
ΔREDAR-V	0.2761	0.4719	0.6892	0.7914	13.30m
ΔREDAR-U	0.2913	0.5025	0.6775	0.7745	16.86m

Table 2: Performances of video recommendation for cold-start users and videos. ΔREDAR-V stands for recommendation of new videos, and ΔREDAR-U stands for recommendation for new users.

The results are shown in Table 2. The ΔREDAR can deal with new users and new videos effectively. Although, undoubtedly, REDAR performs better than ΔREDAR, the degree of degradation of ΔREDAR from REDAR is acceptable. Further, by comparing Table 2 and Table 1, we can see that the performances of ΔREDAR, which does not use any user-video interaction, information of these new users and new videos is still better than other baselines that use this information. Moreover, the ΔREDAR can also be used for online processing of new data in an efficient way. In Table 2, the running time of the incremental processing method ΔREDAR is much less than that of the offline recommendation REDAR for the same number of new users and new videos. The running time is reduced from hours to minutes.

6. CONCLUSION

In this paper, we significantly improved the performance and interpretability of video recommendation systems by learning a common attribute-based representation for users and videos in social media with good interpretability, stability, and appropriate granularity. In order to address the critical challenges including the balance of social attributes and content attributes, the sparsity problem of user-video interactions, and the cold-start problem, we propose our REDAR method based on matrix factorization, in which social attributes and content attributes are flexibly combined, while social and content information is effectively exploited to alleviate the sparsity problem. An incremental version of REDAR is also designed to solve the cold-start problem. The experimental results show that, in most cases, the proposed method can achieve more than 20% relative improvement than state-of-the-art baseline methods.

7. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China, No. 61370022, No. 61303075 and No. 61210008; International Science and Technology Cooperation Program of China, No. 2013DFG12870; National Program on Key Basic Research Project, No. 2011CB302206. Thanks for the support of NExT Research Center funded by MDA, Singapore, under the research grant, WBS:R-252-300-001-490.

8. REFERENCES

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] M. Bertini, A. Del Bimbo, A. Ferracani, F. Gelli, D. Maddaluno, and D. Pezzatini. A novel framework for collaborative video recommendation, interest discovery and friendship suggestion based on semantic profiling. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 451–452. ACM, 2013.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM, 2007.

[5] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM, 2010.

[6] K. El-Arini, M. Xu, E. B. Fox, and C. Guestrin. Representing documents through their readers. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 14–22. ACM, 2013.

[7] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *Computer Vision–ECCV 2012*, pages 530–543. Springer, 2012.

[8] M. Jamali and M. Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 397–406. ACM, 2009.

[9] M. Jiang, P. Cui, R. Liu, Q. Yang, F. Wang, W. Zhu, and S. Yang. Social contextual recommendation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 45–54. ACM, 2012.

[10] M. Jiang, P. Cui, F. Wang, Q. Yang, W. Zhu, and S. Yang. Social recommendation across multiple relational domains. In *ACM CIKM*. ACM, 2012.

[11] M. Jiang, P. Cui, F. Wang, W. Zhu, and S. Yang. Scalable recommendation with social contextual information. *IEEE Transactions on Knowledge Discovery and Engineering*, 2014.

[12] S. Jiang, Y. Zhao, S. Wei, R. Ni, and Z. Zhu. Frame filtering and path verification for improving video copy detection. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 34–37. ACM, 2013.

[13] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[14] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Advances in Knowledge Discovery and Data Mining*, pages 380–389. Springer, 2006.

[15] H. Luo, J. Fan, D. A. Keim, and S. Satoh. Personalized news video recommendation. In *Advances in Multimedia Modeling*, pages 459–471. Springer, 2009.

[16] X. Ma, H. Wang, H. Li, J. Liu, and H. Jiang. Exploring sharing patterns for video recommendation on youtube-like social media. *Multimedia Systems*, pages 1–17, 2013.

[17] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542. ACM, 2002.

[18] T. Mei, B. Yang, X.-S. Hua, L. Yang, S.-Q. Yang, and S. Li. Videoreach: an online video recommendation system. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 767–768. ACM, 2007.

[19] J. Park, S.-J. Lee, S.-J. Lee, K. Kim, B.-S. Chung, and Y.-K. Lee. Online video recommendation through tag-cloud aggregation. *IEEE Multimedia*, 18(1):78–87, 2011.

[20] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[21] K. Stefanidis, E. Pitoura, and P. Vassiliadis. Managing contextual preferences. *Information Systems*, 36(8):1158–1180, 2011.

[22] J. Wang, A. P. De Vries, and M. J. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 501–508. ACM, 2006.

[23] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu. Joint social and content recommendation for user-generated videos in online social network. *IEEE Transactions on Multimedia*, 15(3):698–709, 2013.

[24] S. Xu, H. Jiang, and F. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 83–90. ACM, 2008.

[25] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80. ACM, 2007.

[26] J. C. Yang, Y. T. Huang, C. C. Tsai, C. I. Chung, and Y. C. Wu. An automatic multimedia content summarization system for video recommendation. *Journal of Educational Technology & Society*, 12(1), 2009.

[27] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng. Understanding user behavior in large-scale video-on-demand systems. In *ACM SIGOPS Operating Systems Review*, volume 40, pages 333–344. ACM, 2006.

[28] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 33–42. ACM, 2013.

[29] R. Zhou, S. Khemmarat, and L. Gao. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 404–410. ACM, 2010.

[30] Q. Zhu, M.-L. Shyu, and H. Wang. Videotopic: Content-based video recommendation using a topic model. In *Multimedia (ISM), 2013 IEEE International Symposium on*, pages 219–222. IEEE, 2013.