Video eCommerce: Towards Online Video Advertising

Zhi-Qi Cheng[1,2]*, Yang Liu[2], Xiao Wu[1,#], Xian-Sheng Hua[2,#]
[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
[2]Alibaba Group, Hangzhou, China
zhiqicheng@gmail.com; wuxiaohk@swjtu.edu.cn; {panjun.ly, xiansheng.hxs}@alibaba-inc.com

ABSTRACT

The prevalence of online videos provides an opportunity for e-commerce companies to exhibit their product ads in videos by recommendation. In this paper, we propose an advertising system named *Video eCommerce* to exhibit appropriate product ads to particular users at proper time stamps of videos, which takes into account video semantics, user shopping preference and viewing behavior feedback by a two-level strategy. At the first level, Co-Relation Regression (CRR) model is novelly proposed to construct the semantic association between keyframes and products. Heterogeneous information network (HIN) is adopted to build the user shopping preference from two different e-commerce platforms, Tmall and MagicBox, which alleviates the problems of data sparsity and cold start. In addition, Video Scene Importance Model (VSIM) utilizes the viewing behavior of users to embed ads at the most attractive position within the video stream. At the second level, taking the results of CRR, HIN and VSIM as the input, Heterogeneous Relation Matrix Factorization (HRMF) is applied for product advertising. Extensive evaluation on a variety of online videos from Tmall MagicBox demonstrates that *Video eCommerce* achieves promising performance, which significantly outperforms the state-of-the-art advertising methods.

Keywords

Online Advertising; E-commerce; Video Analysis

1. INTRODUCTION

The online video streaming service business shows the vast market potential. It is forecasted that the online video market in China will post total revenue worth US$ 14.67 billion by 2018, up from 3.75 billion in 2014. In addition, it is reported that the online spending in China will reach one trillion dollars by 2019[1]. The combination of e-commerce platforms and video content providers is inevitable and mutually beneficial, since e-commerce platforms expose their products to videos to increase Gross Merchandise Volume(GMV), and video content providers look forward to traffic monetizing by product exhibition. Unfortunately, video content based product exhibition is still one of the most under-utilized e-commerce strategies, which has an extraordinarily positive impact on consumers.

Traditional personalized product recommendation systems [7, 10, 31] can match consumers with appropriate products by analyzing the user behavior history and video textual tags. Previous browsing and shopping histories may indicate the level of satisfaction with particular products, providing cues for their interests and tastes, which is very useful to enhance user's shopping experience and satisfaction. E-

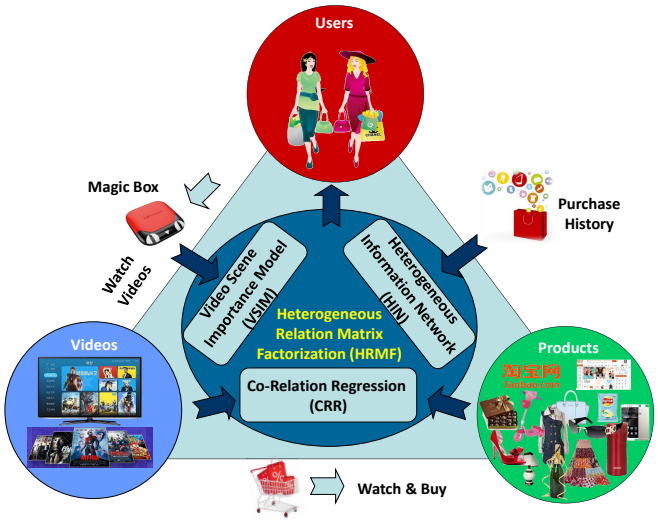

Figure 1: The architecture of Video eCommerce. The associations between videos and products, users and products, as well as users and videos are modeled by Co-Relation Regression (CRR), Heterogeneous Information Network (HIN), and Video Scene Importance Model (VSIM), respectively. With these three components, the personalized video advertising is achieved under Heterogeneous Relation Matrix Factorization (HRMF).

*This work is done while Zhi-Qi Cheng was an intern at Alibaba Group. # indicates corresponding authors.

MM '16, October 15-19, 2016, Amsterdam, Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2964326

[1]https://techcrunch.com/2015/02/04/china-1trillion-ecommerce

commerce companies like Amazon.com and video content providers such as Netflix have made the recommended systems salient parts of their websites. However, few works pay attention to the online video advertising with the combination of recommendation and video content analysis.

Existing video advertising and recommendation methods also encounter several challenges, although visual information has been taken into account. First, recommendation results for most advertising systems are only from searching visually similar products by object detection and image classification [15]. Unfortunately, the category number of pre-trained detectors and classifiers is limited, which results in the poor product coverage. Second, the diversity and abundance of products in e-commerce platforms lead to the association between users and product ads extremely sparse. Especially, since the purchase records of products exhibited on videos are pretty limited, current recommendation systems are always suffered from typical problems, such as cold-start, data sparsity and so on. Third, without considering user preference, state-of-the-art systems recommend the same product ads to all users, which are non-personalized. Fourth, few works have paid attention to the insert positions of product ads to attract users.

In this paper, a novel personalized product advertising system is proposed to recommend product ads to users of Tmall MagicBox. The videos are provided by set-top box, Tmall MagicBox. It is an integrated household digital entertainment system designed by Alibaba, which offers a wide variety of content, including movies, TV series, entertainment programs, documentaries, kids' programs, games and e-commerce. It has over 8,000 movies, 2,600 TV series, 2,100 entertainment programs, and 25,000 educational programs. The products are from Alibaba B2C and C2C online retail platforms, Taobao.com and Tmall.com. The number of products is approximately one billion. In addition, the user preference comes from two different domains, Tmall and Tmall MagicBox. The goal of this work is to transfer shopping behavior of users from e-commerce websites to online video advertising. The architecture of Video eCommerce is illustrated in Figure 1. The associations between videos and products, users and products, as well as users and videos are modeled by Co-Relation Regression (CRR), Heterogeneous Information Network (HIN), and Video Scene Importance Model (VSIM), respectively. With these three components, the personalized video advertising is achieved under Heterogeneous Relation Matrix Factorization (HRMF). To the best of our knowledge, this is the first video e-commerce system has been running on online dataset and comprehensively taking into account the relationships among users, videos and products. Its target is to recommend appropriate product ads to particular customers at proper time stamps, when users are watching movies or TV series. The contributions of this work are as follows:

- Co-relation Regression (CRR) is proposed to model the association between video keyframes and products, by exploiting vast amount of hidden linkages within keyframes, products, and co-relation of keyframes and products.

- Heterogeneous Information Network (HIN) is applied to build the relationship between users and products. Cross-domain user preference is propagated along different meta-paths in HIN to generate latent features for users and products. Therefore, HIN is able to alleviate the problems of data sparsity and cold-start.

- Video Scene Importance Model (VSIM) is proposed to leverage user click-through data and a voting approach is used to automatically adjust the importance of video scenes. In this way, user click-through can be tracked and ads will be embedded at the most attractive positions within the video stream.

- Heterogeneous Relation Matrix Factorization (HRMF) is presented to take the outputs of CRR, HIN and VSIM as the input and mine the relationship among users, keyframes and products. It determines the insert positions in videos of product ads according to user-keyframe relationship, and enables the personalized advertising based on user-product relationship.

This paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 introduces the framework of the proposed advertising system and the preprocessing. Section 4 elaborates the association modeling, including CRR, HIN and VSIM. Heterogeneous Relation Matrix Factorization is presented in Section 5. Experimental results and performance comparison are described in Section 6. Finally, this paper is concluded with a summary.

2. RELATED WORK

This work is closely related to online advertising and product recommendation, which will be briefly reviewed in this section.

2.1 Online Advertising

In the last two decades, there has been substantial volume of works done on online advertising. According to the order of evolution, it can be divided into: (a) content agnostic advertising (random placing advertisements), (b) contextual advertising (placing relevant advertisements based on keywords) [15, 16], and more recently (c) semantic advertising (placing advertisements based on the semantic analysis of the text) [20].

The main objectives of the online advertising are in revenue management through ad allocation. Text-based contextual advertising methods are popular, e.g. Google's AdSense, whose ad allocation ways are optimized by using linear programming [2, 24] or dynamic programming [1, 9, 17]. Although contextual advertising is popular in text-based advertising, contextual multimedia advertising is developed slowly. The trend of online multimedia advertising is summarized and a broad survey on the methodologies for advertising is conducted in [13]. In online image advertising, a contextual advertising system called ImageSense is studied in [14, 16], which can automatically associate relevant ads with an image in the non-intrusive areas. Meanwhile, in online video advertising, a VideoSense system is introduced in [15] which aims to embed more contextually relevant ads at less intrusive positions within the video stream. More recently, the linear in-stream ad allocation problem in an online video is studied in [8] through a dynamic programming approach that accounts for ad prices, ad quality, and externalities from other advertisements. Unfortunately, existing online multimedia advertising algorithms do not consider the personalized user preference.

2.2 Product Recommendation

Recommender systems have been extensively studied in the past decade, and been applied to many successful online services, such as product recommendation at Amazon, movie recommendation at Netflix, video recommendation at Hulu, and music recommendation at Pandora. It has drawn much attention in computer vision, data mining, information retrieval and multimedia communities. Methods based on collaborative filtering [3, 6, 10, 22] have been actively studied recently, which represent the user-item rating matrix with low dimensional latent vectors. A REgularized DualfActor Regression (REDAR) method based on collaborative filtering is proposed in [3], in which social attributes and content attributes are flexibly combined. Although collaborative filtering achieves a huge success, there are two main drawbacks: data sparsity and cold-start. Recently, a large number of works address these two problems directly or indirectly by incorporating additional information, such as social data [11], user feedback [18] and latent factor models [30].

Since the relationship between products is an essential factor for recommendation task, some works try to supply good shopping experience for users by considering product relationship. A novel problem called the bundle recommendation problem is introduced in [32], which takes into account of the dependency of items in the same set. A method is developed in [12] to infer networks of substitutable and complementary products. It is formulated as a supervised link prediction task, where the semantics of substitutes and complements can be learned from data associated with products.

There are also some works for recommendation task based on other factors, such as time, price and so on. The conception of life stage is introduced into product recommendation [7]. The current life-stage of a user is first predicted and corresponding products are then recommended. A stock aware recommender system [31] is developed for Tmall to select recommended items based on both the user preference and the inventory size of items. Although extensive research has been conducted for recommended systems, for video advertising, froementioned research mainly considers the relationship between users and products, the semantics of product and video are totally ignored, which motivates this work.

3. FRAMEWORK AND PREPROCESSING

3.1 Framework

The framework of the proposed online video advertising system, *Video eCommerce*, is illustrated in Figure 2. Preprocessing is first conducted. After video structure analysis, a video is represented as a set of keyframes. Object detection and scene classification are applied on keyframes to obtain the visual-terms (category tags). Different relationships among keyframes, visual-terms, products, product-terms (product textual tags) form five initial association matrices, which are exploited for Co-Relation Regression (CRR). The semantic association between keyframes and products is constructed by CRR. To alleviate the problem of data sparsity and mine user behavior, cross-domain purchase histories from MagicBox and Tmall are integrated to infer user shopping preference, which is modeled with Heterogenous Information Network (HIN). Video Scene Importance Model (VSIM) is integrated to determine the product insertion

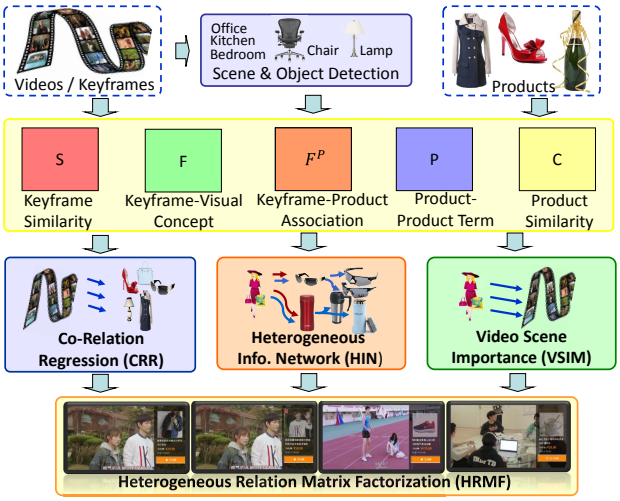

Figure 2: Framework of the proposed *Video eCommerce* advertising system.

position. HRMF is novelly proposed to integrate keyframe-product association, user-product interest relationship and video scene importance to recommend appropriate product ads to users, which is modeled as an optimization problem.

3.2 Object Detection and Scene Classification

After shot boundary detection and keyframe extraction [26], each video is represented as a set of keyframes. A large scale detection network based on deep convolutional models, LSDA framework [5], is adopted to detect objects in keyframes. We fine-tune the layers of 1-7 using our labeled data with bounding box annotation. Totally, the object detection supports 170 object categories that are commonly occurred in videos and have the vast market potentials, such as clothes, furniture, electrical appliances, and so on.

GoogleNet [23] deep convolutional network is applied for scene classification. To accelerate the training process, the size of input images is reduced to 128x128 and small network architecture is used. The stride of conv1 layer is enlarged from 2 to 4, and the kernel number of conv2 layer is reduced from 192 to 96. For remaining Inception layers, the kernel number is set as the half of its original number. Similar to the categories of object detection, the scene categories are determined based on the advertising potentials and appearance frequency in videos, such as office, meeting room, playground, and so on. Totally, 49 scenes are recognized in our work.

3.3 Association Matrix Initialization

The initial association matrices are computed with the assistance of scene classification, object detection and similarity measurement among different components, which are defined as follows:

Modeling the relations between keyframes and visual-concepts (F matrix) The visual-concepts refer to the category labels of object detection and scene classification. The relationship between keyframe f_i and visual-concept t_j is measured based on the $tf-idf$ weight, which is defined as follows:

$$F_{ij} = tf_{ij} \cdot idf_j$$

where tf_{ij} is the term frequency of concept t_j in keyframe

f_i, and idf_j is the inverse document frequency for concept t_j across all keyframes.

Modeling the relations between products and product-terms (P matrix) The product-terms denote the textual description of products. The relationship between product p_i and product-term t_j is defined as follows:

$$P_{ij} = tf_{ij} \cdot idf_j$$

where tf_{ij} is the term frequency of term t_j in product p_i, and idf_j is the inverse document frequency for term t_j across all products.

Modeling of relations between keyframes and products (F^P matrix) The relationship matrix of keyframes and products is denoted as F^P. If a product p_j is relevant to a detected object in keyframe f_i, F^P_{ij} is equal to 1. Otherwise, it is 0. For an object detected in a keyframe, the detected object region is treated as the query example and searched to find visually similar products in database. Meanwhile, products are also searched with text search engine in Tmall using the text label of the object category. These results are further fused as the final relevant products by linear fusion.

Modeling the relations between keyframes (S matrix) The keyframe similarity matrix is denoted as S, in which each item is a linear fusion of visual (V_{ij}) and textual (T_{ij}) similarity between two keyframes.

$$S_{i,j} = e^{-(\alpha V_{ij} + (1-\alpha) T_{ij})}$$

Each keyframe is represented as a 4,096 dimensional CNN feature, and cosine similarity is used to measure their visual similarity V_{ij}. The tf-idf vector in F is adopted to calculate their textual similarity T_{ij} using cosine distance.

Modeling the relations between products (C matrix) Similarly, the product relevance matrix C is a linear fusion of visual and textual similarity between two products, which is the same as S.

4. ASSOCIATION MODELING

The associations between videos and products, users and products, as well as users and videos are modeled by Co-Relation Regression (CRR), Heterogeneous Information Network (HIN), and Video Scene Importance Model (VSIM), respectively, which will be elaborated as following subsections.

4.1 Co-Relation Regression

In order to model the association between keyframes and products, we consider the following two relations from the viewpoints of video semantics and product description, respectively. The relation of video semantics is defined as $E_1 = F^P P$, which assumes that similar keyframes should be mapped to the same product-terms. At the same time, the relation of product description $E_2 = (F^P)^T F$ is also defined, which assumes that similar products should be mapped to the same visual concepts too. With these two co-relations, we have the following three intuitions:

- The co-relations from E_1 and E_2 should be consistent with the initial relations between keyframes and products i.e. F^P.
- If two keyframes have high similarity in S, the association in E_1 should be mapped to the same product-terms.

Algorithm 1 Co-Relation Regression (CRR)

Input: F^P, S, C, F, P, $t_k^{E_1}$, $t_k^{E_2}$, γ_1, γ_2, γ_3.
Output: $\hat{E_1}$, $\hat{E_2}$
1: Initialize: $E_1^0 = F^P P$, $E_2^0 = (F^P)^T F$;
2: **for** $k = 1 \to n$ **do**
3: Calculate $\frac{\partial f_1^k}{\partial E_1}, \frac{\partial f_1^k}{\partial E_2}, \frac{\partial f_2^k}{\partial E_1}, \frac{\partial f_3^k}{\partial E_2}, \frac{\partial g^k}{\partial E_1}$ and $\frac{\partial g^k}{\partial E_2}$
4: $\bigtriangledown F_{E_1}(E_1, E_2) = \frac{\partial f_1^k}{\partial E_1} + \gamma_1 \frac{\partial f_2^k}{\partial E_1} + \gamma_3 \frac{\partial g^k}{\partial E_1}$
5: $\bigtriangledown F_{E_2}(E_1, E_2) = \frac{\partial f_1^k}{\partial E_2} + \gamma_2 \frac{\partial f_3^k}{\partial E_2} + \gamma_3 \frac{\partial g^k}{\partial E_2}$
6: $E_1^{(k+1)} = E_1^k + t_{k+1}^{E_1} \bigtriangledown F_{E_1}(E_1, E_2)$
7: $E_2^{(k+1)} = E_2^k + t_{k+1}^{E_2} \bigtriangledown F_{E_2}(E_1, E_2)$
8: **end for**
9: **return** $\hat{E_1}$, $\hat{E_2}$

- Meanwhile, if two products have high similarity in C, the association in E_2 should be mapped to the same visual concepts.

The issue of keyframe and product association is modeled as an optimization problem. According to these intuitions, an objective function is proposed to solve these two targeting matrices E_1 and E_2. Our target is to minimize the cost function:

$$\begin{aligned} \min_{E_1,E_2} (&\left\| \left(\alpha E_1 P^T + (1-\alpha) F E_2^T\right) - F^P \right\|_F^2 \\ &+\gamma_1 \left\| E_1 E_1^T - S \right\|_F^2 + \gamma_2 \left\| E_2 E_2^T - C \right\|_F^2 \\ &+\gamma_3 (\|E_1\|_F^2 + \|E_2\|_F^2)) \end{aligned} \tag{1}$$

To simplify the description, we get:

- $f_1(E_1, E_2) = \left\| (\alpha E_1 P^T + (1-\alpha) F E_2^T) - F^P \right\|_F^2$ means that the recommended products should not deviate much from the objects detected from the keyframes.
- $f_2(E_1) = \left\| E_1 E_1^T - S \right\|_F^2$ denotes that the recommended products should be similar if two keyframes are semantically similar.
- $f_3(E_2) = \left\| E_2 E_2^T - C \right\|_F^2$ means that two related products should be recommended to the same keyframes.
- $g(E_1, E_2) = \|E_1\|_F^2 + \|E_2\|_F^2$ is a regularization term for smoothing.

where $\alpha \in [0, 1]$ is a weighting factor to control the contribution of keyframes and products, and $\gamma_1, \gamma_2, \gamma_3$ are non-negative parameters less than 1, which control the weights of each constraint.

In order to minimize the cost function, we differentiate $f(E_1, E_2)$ with regard to E_1 and E_2 at each iteration, respectively. The following iterative formulas are proposed for the optimization problem.

$$\frac{\partial f_1}{\partial E_1} = 2\alpha(\alpha E_1 P^T + (1-\alpha) F E_2^T - F^P) P$$

$$\frac{\partial f_1}{\partial E_2} = 2(1-\alpha)((\alpha E_1 P^T + (1-\alpha) F E_2^T - F^P)^T) F$$

$$\frac{\partial f_2}{\partial E_1} = 4(E_1 E_1^T - S) E_1$$

$$\frac{\partial f_3}{\partial E_2} = 4(E_2 E_2^T - C) E_2$$

$$\frac{\partial g}{\partial E_1} = 2\gamma_3 E_1$$

$$\frac{\partial g}{\partial E_2} = 2\gamma_3 E_2$$

The optimization is summarized as an iterative algorithm, which is introduced in Algorithm 1. E_1 and E_2 are updated at each iteration until the function is converged. The results of E_1 and E_2, denoted as $\hat{E_1}$ and $\hat{E_2}$, respectively, are adopted to estimate the keyframe-product correlation by:

$$F^p = \alpha \hat{E_1} P^T + (1-\alpha) F \hat{E_2^T}$$

Time complexity analysis. The time complexity of CRR is determined by the matrix calculation and the number of iterations, which is $O(N_I(N_K+N_P)^2)$, where N_I, N_K, N_P are the numbers of iterations, keyframes and products, respectively.

4.2 Cross-domain User Preference Diffusion

In this section, we will introduce the modeling of user preference, in which user-product association is constructed through meth-paths under the framework of Heterogeneous Information Network (HIN) [28]. The *Video eCommerce* system can be regarded as a HIN, which contains different types of relationships among users and products, as shown in Figure 3. In HIN, two entities (such as, users, videos or products) can be connected via different paths (relations). These paths may contain different entity types and relationship types in inconsistent orders and with various lengths. The meta-path [22] is adopted to describe path types. Previous studies suggest that meta-paths can be used to facilitate entity similarity and proximity measurement [22, 28].

Meta-Path: A meta-path $MP = A_0 \overset{R_1}{\rightarrow} A_1 \overset{R_2}{\rightarrow} ... \overset{R_k}{\rightarrow} A_k$ is a path in a network schema $G_T = (A, R)$, which defines a new composite relation $R_1 \times R_2 \times ... \times R_k$ between type A_0 and A_k, where $A_i \in A$ and $R_i \in R$ for $i = 0, ...k$. $A_0 = sub(R_1) = sub(MP)$, $A_k = obj(R_k) = obj(MP)$ and $A_i = obj(R_i) = sub(R_{i+1})$ for $i = 1, ..., k-1$, where $sub(\cdot)$ defines the subject of certain relationship, and $obj(\cdot)$ defines the object. In our HIN (as showed in Figure 3), the subject of our $MP_1 \sim MP_4$ is users, and the object is products.

Now the cross-domain user preference can be diffused by a couple of meta-paths. The intuition is that if a user purchased a product from Tmall.com directly or through MagicBox, the user and the product will be linked. Meanwhile, other products that are semantically relevant to the purchased product, can be treated as substitutes or alternative ones. The user will be treated as having a hidden relationship with these products. The four meta-paths showed in Figure 3 are listed as follows.

1. $MP_1 : user \overset{purchase}{\rightarrow} product$
2. $MP_2 : user \overset{purchase}{\rightarrow} product \overset{similar}{\rightarrow} product$
3. $MP_3 : user \overset{view\&buy}{\rightarrow} video \overset{contain}{\rightarrow} keyframe \overset{contain}{\rightarrow} object \overset{associate}{\rightarrow} product$
4. $MP_4 : user \overset{view\&buy}{\rightarrow} video \overset{contain}{\rightarrow} keyframe \overset{contain}{\rightarrow} object \overset{associate}{\rightarrow} product \overset{similar}{\rightarrow} product$

in which the observed user-product preference can be directly obtained from meta-paths of MP_1 and MP_3, the potential reference can be induced from meta-paths of MP_2 and MP_4.

The user preference diffusion score between user s and product t along the k_{th} meta-path is calculated as follows:

$$R_k(s,t) = \frac{2*C(s,t)}{C(s,:)+C(:,t)}$$

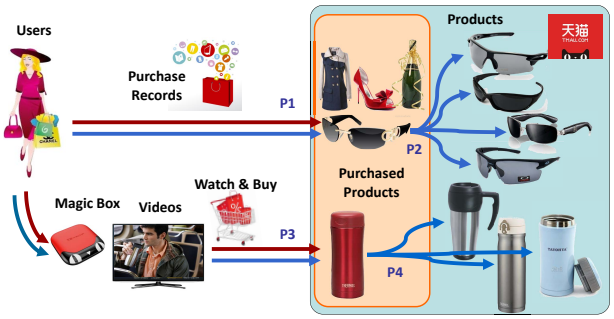

Figure 3: Four meta-paths for cross-domain user preference diffusion between users and products.

where $C(s,t)$ is the number of meta-path instances between s and t. $C(s,:)$ denotes the path count starting with s; and $C(:,t)$ denotes the path count ending with t.

4.3 Video Scene Importance Model

In this section, we will introduce the modeling of video scene importance. As video is a time evolving sequence with diverse contents, users may have different degrees of interest on different parts of the video. The recommended products should be inserted to the keyframes where viewers will stay for a longer time. Therefore, the keyframe is defined as the basic unit of video segment.

In order to obtain the matrix of video scene importance, U^F, we leverage user click-through (user browsing behaviors on a video sequence) to obtain the video scene importance weight. The weight of video scene importance $U^f_{i,j}$ is the importance degree of user u_i to keyframe f_j which is extracted from video shots, where $j = \{1, ..., |D|\}$ and $|D|$ is the number of keyframes in a video. If a user fast-forwards or fast-backwards a scene (i.e., keyframe), he/she may not be interested in this scene, then the weight $U^f_{i,j}$ for keyframe f_j should be decreased. If a user seeks a specific keyframe or to replay a scene, he/she may have strong interest on the content of this scene, then the weight $U^f_{i,j}$ should be increased. Based on these observations, we record the user browsing behaviors and classify the behaviors into four categories. The weight $U^f_{i,j}$ is then dynamically adjusted by a voting-based approach.

The weight $U^{f(t+1)}_{i,j}$ for keyframe f_j at the $t+1$ iteration depends on that in previous step $U^{f(t)}_{i,j}$. It will be added a weight of 0.5 (pause and then browse), 1.0 (seek or replay) and -0.5 (fast browse or skip), respectively, according to user behavior. Please note that $U^{f(t)}_{i,j}$ is normalized to [0, 1] by

$$U^{f(t)}_{i,j} = \frac{U^{f(t)}_{i,j} - U^{f(t)}_{min}}{U^{f(t)}_{max} - U^{f(t)}_{min}}$$

where $U^{f(t)}_{max}$ and $U^{f(t)}_{min}$ denote the maximum and minimum of $U^{f(t)}_{i,j}$ in the t_{th} iteration, respectively. When a viewer watches the video which contains keyframe f_j more than once, $U^{f(t)}_{i,j}$ directly indicates the video scene importance for user u_i.

For a new viewer who has not seen the video before, based on the assumption that watching behavior of reviewers is approximately homophily [4], we use the average of all users who have seen the video before to initialize its weight.

5. HETEROGENEOUS RELATION MATRIX FACTORIZATION

With the association modeling presented in last section, the personalized video advertising is achieved under Heterogeneous Relation Matrix Factorization (HRMF). Three target matrices can be defined for aforementioned three models. The first target matrix is user-keyframe preference matrix $\hat{U_f}$, which denotes the interests between viewers and keyframes. It is utilized to find the best insert position of product ads. The second one is keyframe-product association matrix $\hat{F}_p$, referring to the association between keyframes and products, which combines user preference and video semantics. The third one is meta-path weigh matrix W, denoting the importance of user-product interaction under certain meta-path semantics. It is used to fuse cross-domain user preference elegantly.

The personalized product advertising is converted to an optimization problem. The first term of the model incorporates the collaborative filtering component, which keeps the recommended product results closer to the observed user-product interactions. The second and third items consider the video scene importance and the association between keyframes and products, respectively. The fourth term of the model is the user-product relationship from the cross-domain user preference diffusion method. The last one is the smoothing term. It is defined as follows:

$$J = \min_{\hat{U^f},\hat{F^p},W} (\sum_{i=0}^{N_U}\sum_{j=0}^{N_K}(\hat{U_i^f}\hat{F_j^p} - R_{i,j})^2 + \alpha\sum_{i=0}^{N_U}\|(\hat{U_i^f} - U_i^f)\|_F^2 + \beta\sum_{j=0}^{N_K}\|\hat{F_j^p} - F_j^p\|_F^2 + \mu\sum_{k=0}^{N_M}W_k\sum_{i=0}^{N_U}\sum_{j=0}^{N_K}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}^k)^2 + \lambda(\|\hat{U^f}\|_F^2 + \|\hat{F^p}\|_F^2 + \|W\|_F^2)) \quad (2)$$

These symbols have the same meanings as introduced in previous sections. N_U, N_K, N_M are the number of users, keyframes and meta-paths, respectively. $R_{i,j}$ is the purchase records that user i purchased product j in Tmall MagicBox. α and β are parameters capturing the importance of video scene and video semantics, respectively, which will be discussed in the experimental part. Similar to [19], the logistic function $f(x) = 1/(1 + exp(-x))$ is used to normalize the values in U^f, F^p and $R_{i,j}^k$ within the range of $[0, 1]$.

The learning of HRMF is implemented in a two-step iteration approach, where the preference matrix of viewers to keyframes $\hat{U^f}$, the keyframe-product association matrix $\hat{F^p}$ and the weight matrix for meta-paths W will be mutually enhanced. In the first step, the weight matrix W is fixed and the optimal $\hat{U^f}$ and $\hat{F^p}$ are learnt. In the second step, $\hat{U^f}$ and $\hat{F^p}$ are fixed, and the optimal weight matrix W is learnt. The optimization is summarized as an iterative algorithm, which is introduced in Algorithm 2.

Step 1: Optimize $\hat{U^f}$ and $\hat{F^p}$ given W: When W is fixed, the model becomes a traditional collaborative filtering method. Therefore, similar to [27], Stochastic Gradient Descent (SGD) is adopted to solve this problem.

Step 2: Optimize W given $\hat{U^f}$ and $\hat{F^p}$: When $\hat{U^f}$ and $\hat{F^p}$ are fixed, it only includes $\hat{U^f}$, and $\hat{F^p}$ can be discarded. The objective function is reduced to:

Algorithm 2 Heterogeneous Relation Matrix Factorization (HRMF)

Input: F^p of CRR, R^k of HIN, U^F of VSIM, $R_{i,j}$ of Tmall MagicBox, and learning rate $\alpha^s \in (0, 1)$.
Output: $\hat{R}$

```
Initialize Û^f, F̂^p, W randomly;
while Û^f, F̂^p are not converged do
    for each observed user-product pair (i, j) ∈ R do
        Calculate ∂J/∂Û_i^f, ∂J/∂V̂_j^p,
        Û_i^f ← Û_i^f − α^s ∂J/∂Û_i^f
        V̂_j^p ← V̂_j^p − α^s ∂J/∂V̂_j^p,
    end for
    while W is not converged do
        Calculate ∂J_1/∂W_k,
        W_i ← W_i − α^s ∂J_1/∂W_k,
    end while
end while
return The predicted ratings R̂ = Û^f F̂^p.
```

$$J_1 = \mu\sum_{k=0}^{N_M}W_k\sum_{i=0}^{N_U}\sum_{j=0}^{N_K}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}^k)^2 + \lambda(\|W\|_F^2)$$

We can see that J_1 becomes a linear model for each W_k. Therefore, Stochastic Gradient Descent is also used to obtain W.

$$\frac{1}{2}\frac{\partial J}{\partial \hat{U_i^f}} = \sum_{i=0}^{N_U}\hat{U_i^f}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}) + \alpha\sum_{i=0}^{N_U}(\hat{U_i^f} - U_i^f) + \mu\sum_{k=0}^{N_M}\sum_{i=0}^{N_U}\hat{U_i^f}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}^k) + \sum_{i=0}^{N_U}(\hat{U_i^f})$$

$$\frac{1}{2}\frac{\partial J}{\partial \hat{V_j^p}} = \sum_{j=0}^{N_K}\hat{V_j^p}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}) + \beta\sum_{j=0}^{N_K}(\hat{V_j^p} - V_j^p) + \mu\sum_{k=0}^{N_M}\sum_{j=0}^{N_K}\hat{V_i^f}(\hat{U_i^f}\hat{F_j^p} - R_{i,j}^k) + \sum_{j=0}^{N_K}(\hat{V_j^p})$$

$$\frac{1}{2}\frac{\partial J_1}{\partial W_k} = \lambda W_k + \mu\sum_{i=0}^{N_U}\sum_{j=0}^{N_K}(\hat{U_i^f}\hat{F_j^p} - R_{i,j})^2$$

After minimizing this model, $\hat{U^f}$, $\hat{F^p}$, and W can be obtained. The predicted ratings can be obtained as $\hat{R} = \hat{U^f}\hat{F^p}$. In practice, the recommended insert positions should meet the uniform distribution. Assume that there are M insert positions, the video will be divided into M equal portions. For a user u_i, according to the maximun $\hat{R_{i,j}}$, the product p_j will be recommended at the keyframe which is indicated by maximum $\hat{U_i^f}$ in each portion.

Time complexity analysis. The time complexity of computing the gradients of U, F and W is $O(N_I N_U)$, $O(N_I N_U)$ and $O(N_I N_U N_P)$, respectively. N_I, N_K, N_P and N_U are the number of iterations, keyframes, products and users, respectively. Therefore, the upper bound of the complexity of Alg. 2 is $O(N_I N_U N_P)$

6. EXPERIMENTS

In this section, we evaluate the effect of individual component for the performance of *Video eCommerce*, and compare it with state-of-the-art approaches. We study this problem empirically by conducting a live controlled experiment with real customers in a real industrial setting.

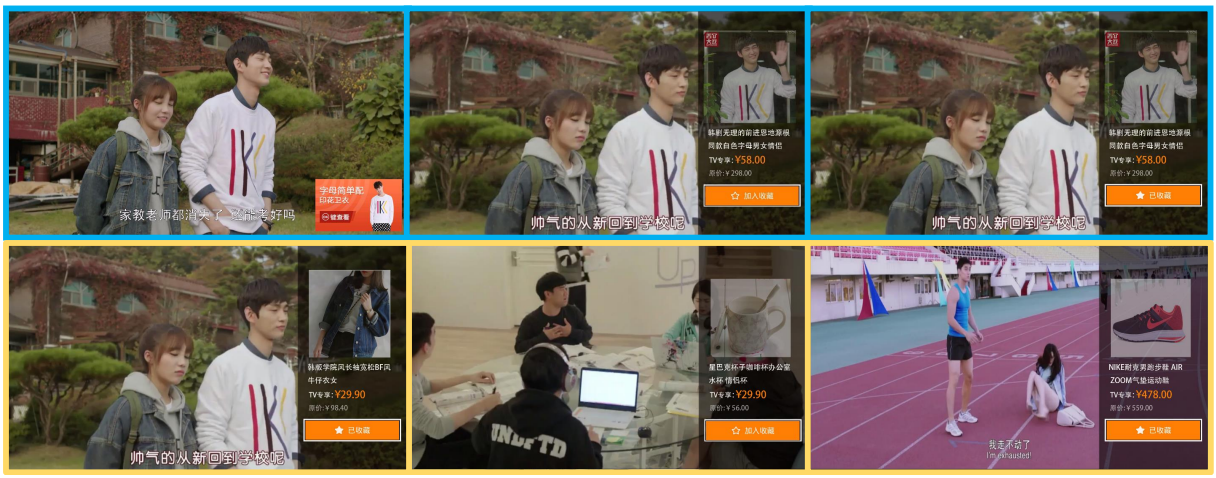

Figure 4: The personalized product advertising examples recommended by *Video eCommerce* for different users. The top of the figure (with blue border) shows a complete advertising procedure.

6.1 Experimental Design

To eliminate the bias from user sampling, we select at least 400 users having purchase records from every video. Several statistical tests, including t-test and Chi-square test, are conducted to keep the consistency between the selected users and whole customer database, based on six variables: 1) age; 2) gender; 3) number of orders; 4) number of purchased items at different domains; 5) money spent on the Website; 6) year of subscription. We measure the distributions of each variable between the selected users and the customer database, and update the selected users until there is no any statistically significant difference on similarities. Each performance metric is then averaged across customers instead of using the absolute values for every video.

To evaluate the performance, we select the latest released 235 videos and their associated information as our dataset, which consists of 25 movies, 4 TV series (157 videos) and 53 variety shows. Totally, there are 47,768 users, 235 videos, 52,405 keyframes, 288,200 products, 219 visual concepts and 10,932 product terms.

6.2 Evaluation Criteria

Video eCommerce is an online system, we focus on business related performance matrices. A complete advertising procedure is illustrated on the top of Figure 4 (with blue border). From left to right, when a user views a video, the *Video eCommerce* will pop up an ads on bottom right corner of the screen. If a user is interested in this product ads, he/she may click ads. By viewing the details, if the user is really interested in it, he/she may favorite it. At this point, the business-related crucial performance measures are defined as follows:

- **Page View Click Ratio (PVC)**: It refers to the ratio of product ads which are clicked to all product ads that are popped up.
- **Unique Visitor Click Ratio (UVC)**: It is the ratio of the unique individuals who click product ads to total individuals who view product ads, regardless how many product ads they click.
- **Page View Favorite Ratio (PVF)**: It denotes the ratio of product ads which are favorited to all product ads that are clicked.
- **Unique Visitor Favorite Ratio (UVF)**: It refers to the ratio of unique individuals who favorite products to total individuals who click product ads, regardless how many products they favorite.

6.3 Performance Comparison of CRR Model

In this subsection, we evaluate the impact of CRR to the *Video eCommerce* system.

6.3.1 *Baselines*

To evaluate the performance, we compare the proposed method with three baseline methods (Search_T, Search_V and Search_T+V) and four variants of CRR. The results of these methods are treated as the input of HRMF model to evaluate the overall performance. Search_T, Search_V and Search_T+V represent that only textual features (the tags of object and scene detection), only visual features (CNN features of detected objects), and their combinations are used to retrieve the similar products, respectively. Meanwhile, to evaluate the effects of different components of CRR, four variants of CRR are also tested. CRR-KA and CRR-PA refer to CRR with keyframe relation and product relation is considered, when α is set to 1 and 0 in Eqn. 1, respectively. CRR-KR and CRR-PR refer to CRR with regularizer of keyframes and products when γ_1 or γ_2 is set to 0 in Eqn. 1, respectively.

6.3.2 *Experiment Results*

The performance comparison of CRR model according to different categories is shown in Figure 5. Generally, the performance of Search_T is poor, since products with the same text words are too monotonous, so that recommendation purely based on text cannot meet the requirements of users. When users are attracted by a video, visually similar products are more accepted for users. Therefore, Search_V association has a better performance than Search_T. The combination of textual and visual features improves the performance, since users will meet a lot of bizarre ads when sys-

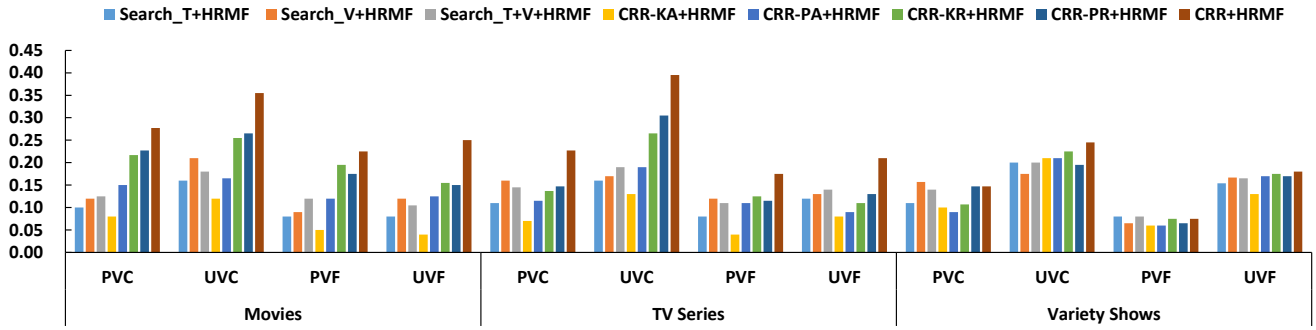

Figure 5: Performance comparison for CRR model.

Table 1: Performance comparison for HIN and VSIM models

	Movies				TV Series				Variety Shows			
Approach	PVC	UVC	PVF	UVF	PVC	UVC	PVF	UVF	PVC	UVC	PVF	UVF
Tmall	0.18	0.20	0.16	0.14	0.18	0.25	0.16	0.12	0.13	0.19	**0.08**	0.13
MagicBox	0.16	0.16	0.13	0.13	0.12	0.20	0.10	0.10	0.11	0.16	0.06	0.10
HIN-Equal	0.22	0.26	0.20	0.16	0.16	0.29	0.15	0.13	0.15	0.23	**0.08**	**0.18**
HIN-Random	0.22	0.27	0.18	0.16	0.16	0.32	0.13	0.13	0.14	0.19	0.07	0.17
HIN	**0.28**	**0.36**	**0.23**	**0.25**	**0.23**	**0.40**	**0.18**	**0.21**	0.15	**0.25**	**0.08**	**0.18**
VSIM-Un	0.16	0.17	0.16	0.14	0.13	0.21	0.13	0.11	0.13	0.21	0.06	0.17
VSIM-Random	0.24	0.26	0.16	0.16	0.16	0.28	0.14	0.14	**0.16**	0.20	**0.08**	0.17
VSIM	**0.28**	**0.36**	**0.23**	**0.25**	**0.23**	**0.40**	**0.18**	**0.21**	0.15	**0.25**	**0.08**	**0.18**

tem only relies on the visual information. And because users have dissimilar viewing habits for different video categories, which contain completely diverse semantic information, the performance for different categories are inconsistent. Furthermore, most of variety show videos in our dataset are talk show, which contain little semantic information, so the improvement for variety shows is minor.

CRR-KA has the worst performance than CRR-PA, since the semantic of video is more important than product when the user is concentrate on videos. All regularizers imposed to the main regression term play important roles in CRR model. Comparatively, the keyframe regularizer (CRR-KR) is more helpful than the product regularizer (CRR-PR). It indicates that video semantics are more useful than the relationship between products for *Video eCommerce*. Finally, it is easy to find that CRR model well addresses the association between keyframes and products, which achieves the best performance.

6.4 Performance Comparison of HIN Model

To evaluate the performance of the proposed HIN model, in this subsection, we analyze the impact of cross-domain user shopping preference.

6.4.1 *Baselines*

We compare the proposed method with single-domain baselines Tmall and MagicBox. That is, only MP1 and MP2 meta-paths are used for Tmall, while MP3 and MP4 meta-paths are deployed for MagicBox. In addition, we also evaluate the performance of different weighs for meta-paths. HIN-Equal and HIN-Random mean that weights for each selected meta-path are equal and randomly selected, respectively. For HIN, the weights (learned by HRMF) of four meta-paths are assigned as 0.40, 0.24, 0.21 and 0.15, respectively.

6.4.2 *Experiment Results*

The performance comparison is listed in Table 1. We can see that the performance using Tmall purchase history has better performance than MagicBox. This is because the number of purchase records from MagicBox is much sparser than Tmall, since MagicBox is still a newborn product. In this situation, the user preference induced from Tmall is very meaningful. Furthermore, HIN has better performance than HIN-Equal and HIN-Random. This is because our HRMF model can automatically learn regression weights, and do not need to set parameters in advance. HIN model not only integrates clues from Tmall and MagicBox, but also learns suitable weights to integrate user performance, which outperforms the single-domain methods and other HIN variants.

6.5 Performance Comparison of VSIM Model

In this subsection, the impact of video scene importance is conducted.

6.5.1 *Baselines*

To evaluate the impact of the proposed VSIM model, we compare it with two variants of VSIM, VSIM-Un and VSIM-Random, which mean that video scene importance is not considered, and it is randomly selected, respectively.

6.5.2 *Experiment Results*

The experimental results are reported in Table 1. We can see that the performance of VSIM-Un is poor, when video scene importance is ignored. VSIM and VSIM-Random have conspicuously better performance than VSIM-Un. It means that user attention based on video scene importance can be rationally used to improve the performance. VSIM model outperforms VSIM-Random, indicating that VSIM model can choose suitable video scene weights to insert advertisement.

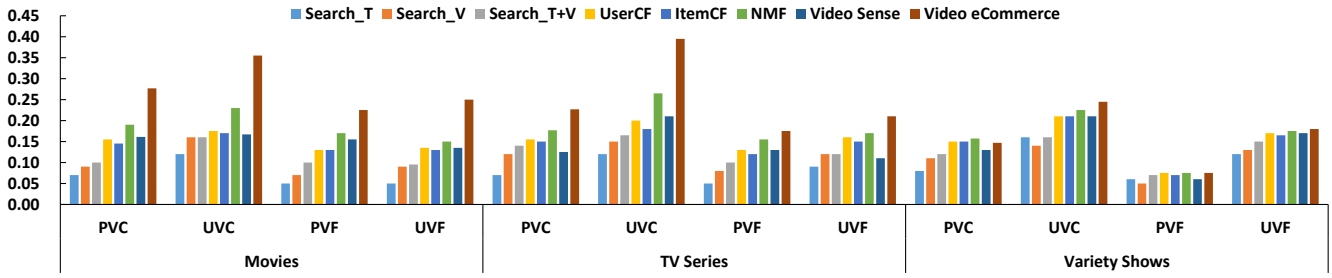

Figure 6: Performance comparison of the state-of-the-art methods

6.6 Performance Comparison With State-of-the-art Approaches

In this subsection, we will compare the performance of *Video eCommerce* with state-of-the-art approaches.

6.6.1 Baselines

To verify the performance, we compare our system with non-personalized methods, such as retrieval-based methods (Search_T, Search_V and Search_T+V) and VideoSense [15]. Here, Search_T, Search_V and Search_T+V are different from Sec. 6.3 where HRMF is not applied. We also compare it with three commonly used recommender systems, UserCF [25], ItemCF [21], and NMF [29]. UserCF [25] and ItemCF [21] are the most popular user-based and item-based collaborative filtering approaches, respectively. NMF [29] is a non-negative matrix factorization method. For these recommender approaches, the results from CRR are used to associate videos to corresponding products.

6.6.2 Experiment Results

The performance comparison is shown in Figure 6. Generally, the personalized recommender systems have much better performance than non-personalized ones, such as traditional retrieval-based methods. This is because the click and favorite behaviors of users are more related to their personalized interests. For classic recommendation methods, NMF performs much better than UserCF and ItemCF, since it has better performance in approximating the user-product interactions. People may have an impulse to buy semantically related products, so VideoSense has better performance than traditional retrieval-based methods. Video eCommerce combines video semantics and user preference, which performs better than VideoSense. Compared with these personalized recommendation, Video eCommerce achieves the best performance across all measures and obtains approximately 25% improvement in terms of PVC compared to the best baseline NMF. Furthermore, the experiments demonstrate that video content (semantics, scene, and association rule) and cross-domain user shopping preference have direct effects on the performance of video recommendation. The proposed *Video eCommerce* is an effective recommendation system, which can recommend appropriate products by simultaneously considering product property, video semantics and user's shopping preference. It can meet user's requirement and bring significant value for e-commerce websites.

6.6.3 Advertising Examples

Video eCommerce can recommend different products to various users according to their preference, even if they are watching the same keyframe. As a user has no obvious shopping preference, visually similar products will be advertised. As shown in the bottom of Figure 4, when a user has a clear preference (e.g., she likes jeans), a jeans is recommended. What's more, Video eCommerce also considers video semantics. A coffee cup and Nike running shoe are advertised according to video semantics, in which these two objects are not detected in the keyframes.

6.7 Parameter Study

There are four parameters in CRR and four parameters in HRMF. For the parameter setting, grid search is used to obtain the optimal parameters. To verify the importance of co-relation, we test the performance of α from 0 to 1 with the step size of 0.1. In our application, the best result is achieved when $\alpha = 0.6$. As the value of α is too small, the largest performance degradation is occurred, referring to CRR-KA in Figure 5. When it is too large, the result will not be good enough, as CRR-PA in Figure 5.

In HRMF, to verify the contextual and video scene constraint, we test the performance of α and β from 0 to 20 with the step size of 0.1. In our application, the best result is achieved when $\alpha = 10$ and $\beta = 4$. When the values of α and β are too small, the largest degree of performance degradation is occurred. From the experimental experience, μ is an important parameter, which directly affects the performance of our algorithm. When R is sparse, a larger μ can improve the recommendation results, because more information can be added to the training process. On the other hand, when R is not sparse, a larger μ will bias the recommendation results. Therefore, the value of μ depends on how sparse R is. In this sense, we can use the proportion of non-zero elements in matrix R to calculate μ. In our application, the best result is achieved when $\mu = 0.75$.

7. CONCLUSIONS

In this paper, an innovative system, *Video eCommerce*, is presented for online video advertising, which is able to exhibit appropriate product ads to particular users according to video content. Co-Relation Regression, Heterogeneous Information Network, Video Scene Importance are proposed to effectively model the associations among users, videos and products, which ensure the product diversity, alleviate the problems of data-sparsity and cold-start, and portray the importance of video content. Finally Heterogeneous Relation Matrix Factorization is applied for product recommendation. Extensive experiments have been conducted over a large-scale of online video dataset and promising results have been achieved in crucial business-related performance measures, which outperforms state-of-the-art approaches. To the best of our knowledge, this is one of the first video e-commerce systems have been running online comprehensive-

ly taking into account the relationships among users, videos and products. In the future, we will explore large scale online advertising and conduct real-time update for new objects, so that the whole framework can be efficiently performed in real time.

8. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61373121), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (Grant No. 13QNJJ0149), and Graduate Innovation Fund of Southwest Jiaotong University (Grant No. YC201504110).

9. REFERENCES

[1] G. Aggarwal, J. Feldman, S. Muthukrishnan, and M. Pál. Sponsored search auctions with markovian users. In *Internet and Network Economics*, pages 621–628. 2008.

[2] D. M. Chickering and D. Heckerman. Targeted advertising on the web with inventory management. *Interfaces*, 33(5):71–77, 2003.

[3] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *ACM MM*, pages 597–606, 2014.

[4] K. Eyal and A. M. Rubin. Viewer aggression and homophily, identification, and parasocial relationships with television characters. *Journal of Broadcasting & Electronic Media*, 47(1):77–98, 2003.

[5] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *NIPS*, pages 3536–3544, 2014.

[6] S. Isaacman, S. Ioannidis, A. Chaintreau, and M. Martonosi. Distributed rating prediction in user generated content streams. In *RecSys*, pages 69–76, 2011.

[7] P. Jiang, Y. Zhu, Y. Zhang, and Q. Yuan. Life-stage prediction for product recommendation in e-commerce. In *KDD*, pages 1879–1888, 2015.

[8] W. Kar, V. Swaminathan, and P. Albuquerque. Selection and ordering of linear online video ads. In *RecSys*, pages 203–210, 2015.

[9] D. Kempe and M. Mahdian. A cascade model for externalities in sponsored search. In *Internet and Network Economics*, pages 585–596. 2008.

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

[11] H. Ma. On measuring social friend interest similarities in recommender systems. In *SIGIR*, pages 465–474, 2014.

[12] J. J. McAuley, R. Pandey, and J. Leskovec. Inferring networks of substitutable and complementary products. In *KDD*, pages 785–794, 2015.

[13] T. Mei and X.-S. Hua. Contextual internet multimedia advertising. *Proceedings of the IEEE*, 98(8):1416–1433, 2010.

[14] T. Mei, X.-S. Hua, and S. Li. Contextual in-image advertising. In *ACM MM*, pages 439–448, 2008.

[15] T. Mei, X.-S. Hua, and S. Li. Videosense: A contextual in-video advertising system. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(12):1866–1879, 2009.

[16] T. Mei, L. Li, X.-S. Hua, and S. Li. Imagesense: towards contextual image advertising. *ACM TOMM*, 8(1):6, 2012.

[17] G. Roels and K. Fridgeirsdottir. Dynamic revenue management for online display advertising. *Journal of Revenue & Pricing Management*, 8(5):452–466, 2009.

[18] I. Ronen, I. Guy, E. Kravi, and M. Barnea. Recommending social media content to community owners. In *SIGIR*, pages 243–252, 2014.

[19] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2007.

[20] S. Sarawagi. Information extraction. *Foundations and trends in databases*, 1(3):261–377, 2008.

[21] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, pages 285–295, 2001.

[22] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *PVLDB*, 4(11):992–1003, 2011.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv:1409.4842*, 2014.

[24] J. Turner, A. Scheller-Wolf, and S. Tayur. Or practice-scheduling of dynamic in-game advertising. *Operations research*, 59(1):1–16, 2011.

[25] J. Wang, A. P. De Vries, and M. J. T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *SIGIR*, pages 501–508, 2006.

[26] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *ACM MM*, pages 218–227, 2007.

[27] B. Yang, Y. Lei, D. Liu, and J. Liu. Social collaborative filtering by trust. In *AAAI*, pages 2747–2753, 2013.

[28] X. Yu, X. Ren, Y. Sun, Q. Gu, B. Sturt, U. Khandelwal, B. Norick, and J. Han. Personalized entity recommendation: A heterogeneous information network approach. In *WSDM*, pages 283–292, 2014.

[29] S. Zhang, W. Wang, J. Ford, and F. Makedon. Learning from incomplete ratings using non-negative matrix factorization. In *SDM*, volume 6, pages 548–552, 2006.

[30] Y. Zhang, M. Zhang, Y. Zhang, G. Lai, Y. Liu, H. Zhang, and S. Ma. Daily-aware personalized recommendation based on feature-level time series analysis. In *WWW*, pages 1373–1383, 2015.

[31] W. Zhong, R. Jin, C. Yang, X. Yan, Q. Zhang, and Q. Li. Stock constrained recommendation in tmall. In *KDD*, pages 2287–2296, 2015.

[32] T. Zhu, P. Harrington, J. Li, and L. Tang. Bundle recommendation in ecommerce. In *SIGIR*, pages 657–666, 2014.