IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2018, accepted February 27, 2018, date of publication March 2, 2018, date of current version April 18, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2811463

Personal Tastes vs. Fashion Trends: Predicting Ratings Based on Visual Appearances and Reviews

YINING LIU AND YANMING SHEN
Dalian University of Technology, Dalian 116024, China

Corresponding author: Yining Liu (lyn.workmail@gmail.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant DUT17ZD303, in part by the National Key Research and Development Program of China under Grant 2016YFB1000205, in part by the State Key Program of National Natural Science of China under Grant 61432002, and in part by the Dalian High-level Talent Innovation Program under Grant 2015R049.

ABSTRACT People have their own tastes on visual appearances of products from various categories. For many of them, the tastes are affected by the current fashion trend. Studying visual appearances and fashion trend makes us understand the composition of users' preferences and their purchase choices. However, since the fashion is changing over time, it is complex to model time-aware and non-time-aware variables simultaneously. In this paper, we present VIsually-aware Temporal rAting modeL with topics using review text to help mine visual dynamics and non-visual features for rating prediction task. Understanding the reviews will help the Recommender Systems (RSs) know whether a user is attracted by the appearance of an item, and which aspect of an item's appearance contributes most to its ratings. To achieve this, we incorporate the visual information into the rating predicting function and introduce a topic model that can automatically classify words in an item's reviews into non-visual words that explain the coherent feature, and visual words that are associated with its visual appearances in each time period, respectively. We run experiments on eleven real-world public datasets and the results show that our model performs better on predicting ratings than many of the state-of-the-art RSs, such as PMF, *time*SVD++, HFT, JMARS, ETDR, and TVBPR+.

INDEX TERMS Recommender system, collaborative filtering, rating prediction, visual-aware, topic model, fashion-aware, temporal dynamics.

I. INTRODUCTION

Recommender Systems(RSs) play a critical role in helping users discover and evaluate products and services by modeling the complex preferences that people exhibit toward items based on their past interactions and feedback [1], [2]. To achieve better expressive power, some previous works have made use of information such as temporal dynamics, review text or the content of the items themselves [3]–[6]. In recent years, researchers found that visual appearances have a great influence on users' choices in many applications such as clothing and artistic recommendations [7]–[11]. In such scenarios, visual dynamics play an important role in helping users make purchase decisions. In another word, item visual appearances might contain some features that attract people with certain preferences to buy them. We are interested in exploiting these features and making use of them to give better recommendations. To achieve this, we need not only item visual information but also temporal dynamics and review text. We believe that item visual information will somehow be discussed and reflected in their reviews, and both personal tastes and fashion trends are changing over time.

However, there are several issues that make it difficult to model these factors simultaneously. First, it is simply the scale of the data and the complexity of the factors involved; fully utilizing the acquired information shall require large corpora of products (including ratings, time information, images and review text) and big quantity of variables. Second, it is the fact that visual appearances are only a part of the items' features, which means that the model needs to be capable of differentiating the visual features from the non-visual ones. Third, since the fashion is changing over time, it is complex to model time-aware and non-time-aware variables simultaneously, especially when review text is involved. Finally, it is important to balance individuals' personal tastes and the fashion trends; the amount of information contained in items' visual images differs widely; and people are not affected by the visual images in the same way.

Our main goal is to build a visually-aware temporal rating model that captures user and item latent features based on both item visual appearances and reviews. The popularity of a product is decided not only by its non-visual features such as quality, price and material, but also by the features considered 'fashionable' like the style, color or other aspects of its appearance design. Studying visual appearances makes us understand the composition of users' preferences and their purchase choices. For example, people loving the same colors or having close dressing tastes might potentially have similar views on other aspects. As the fashion evolves, the contribution of the visual appearances to an item's ratings changes. For instance, a shoe might be popular for a time for its fashionable design, but when the fashion changes, some people may still like it because of its other features such as comfortable, leather or light that match their personal tastes. Therefore, the ability of separating these two kinds of features and tracking the fashion evolution process are key for building such a model.

Review text is another valuable information which has been proved to be effective in helping RSs understand users' choices and improving recommendation accuracy [12]–[15]. Reviews generated by users contain rich information about their preferences [16]–[20]. For example, a user might explain in her review the reason she gives a certain rating to an item. The reviews will help RS understand what a user mainly cares about, and which aspect of an item contributes most to its ratings. By jointly modeling visual appearances and review text, we can build a model that automatically finds words in reviews of an item that best explain its visual features, therefore exploiting the latent common features of users who share the same views on the products with similar looks. In the meantime, it can also capture words that express the items non-visual features that differentiate it from others.

In this paper, we develop a new model to address our goals above. Although many previous studies have considered temporal dynamics, visual appearances, or review text, few have modeled all of them simultaneously. The existing time-aware models such as [6] divide the rating data into multiple time periods, and assign each time period a unique set of latent features for each user and item to capture their temporal dynamics. Recently, researchers proposed methods that combine visual appearances with time-aware models to uncover the visual temporal dynamics [21]. However, these visual-aware models are not designed for rating prediction task and all the RSs above do not make use of the rich information contained in user reviews. The review will help a RS understand whether a user is attracted by the appearance of an item, what the user mainly cares about, and which aspect of the item contributes most to its ratings. It is therefore tempting to study how temporal RSs can explore reviews and capture visual dynamics to address the rating sparsity issue and achieve higher accuracy. To achieve this, we decompose the review text of each item into two types of words—*non-visual* words and *visual* words, and design a topic model considering visual appearances and temporal dynamics that is capable of distinguishing the words automatically. Non-visual words explain the coherent features of an item that almost unchanged across time periods. In contrast, visual words refer to those that depict the fashion elements contained in the visual appearances of each item in time periods. As only a modest amount of parameters are affordable per item due to the huge item vocabulary involved, temporal factors are shared by items' visual information in each time period. Thus, we could exploit the fashion trends in each time period as well as the visual words for each item.

A. CONTRIBUTIONS

The main contribution of our work is to mine user and item latent features and the influence of fashion on the ratings from both item visual appearances and reviews. Specifically, we claim the following benefits over the existing approaches.

1) The proposed VITAL model jointly mines the temporal changes in visual factors of items and their latent non-visual features, together with the associated review text in a single learning stage. We believe visual information play an important role in users' purchasing decisions. And this influence changes as the fashion evolves. Modeling it helps us achieve better rating prediction performance.
2) By transferring item latent features to the topic distributions, the proposed VITAL model can learn the word topic distributions for words in the vocabulary, which can be further utilized to automatically tag the products.
3) In experiments with eleven real-world datasets where visual decision factors are at play, our model significantly outperforms the state-of-the-art RSs, such as MF, timeSVD++, HFT, JMARS, ETDR and TVBPR+ on rating prediction accuracy.

B. RELATED WORK

1) MODELING TEMPORAL DYNAMICS

Collaborative filtering techniques have been widely used to mine contextual information embedded in ratings [22]. And time information is considered as one of the most useful context dimensions in the past few years. It facilitates tracking the evolution of user preferences and identifying periodicity in user habits and interests. There exists a multitude of studies utilizing time information to improve the recommendation accuracy. Reference [6] captures user preferences to improve the hit ratio of Top-N recommendations by learning long-term and short-term factors. Previous work in [5] studies the sequence of user behaviors and reveals the relation between user ratings and their expertise. Koren [23] proposed a time-aware RS named timeSVD++ which divides the rating data into time slots, and gives each time slot a separate set of parameters for each user and item to capture the temporary dynamics of their latent features.

2) MODELING REVIEW TEXT

Earlier studies demonstrated the importance of considering review text in RSs [12]–[14]. In this type of models, a topic model is often used to explain the review text. The topic model is then combined with a rating prediction model for predicting ratings. To connect the two models, one must map the users/items' feature vectors in the MF model to the parameters in the topic models. For example, in HFT model [13], the authors transform an item i's feature vector γ_i in MF model to the topic distribution of its review text θ_i in the LDA [24] model with a transformation function. In another proposed model JMARS [12], Diao *et al.* decomposed the user and item feature vectors into many parameters in their MF model, and then designed three transformation functions. They select one of them for every word in review texts based on which component this word belongs to. Reference [25] proposed a recommendation model named TriRank to cast the task as vertex ranking and devised a generic personalized algorithm for ranking in tripartite graphs. To create such graphs, the authors extracted aspects from textual reviews to enrich the user-item binary relations to a user-item-aspect ternary relations. Reference [14] adopted a mixture of Gaussian to model the ratings, but kept using LDA as topic model to learn the latent topics in reviews. Reference [26] proposed a recommendation framework named MR3, which jointly modeled users' rating behaviors, social relationships, and review comments. The experiments show that models combined with review text can achieve better performance on predicting ratings than pure MF models since its ability of finding explanations from reviews for ratings.

3) VISUALLY-AWARE COLLABORATIVE FILTERING

It is beneficial to consider visual information to build recommender systems that are able to understand the visual aspects of the user-item interactions. Jagadeesh *et al.* [8] gathered a street fashion data set of images and build a automated large scale visual RS. In [7], He and McAuley map users and items into a visual space with the inner products depicting the visual compatibility. However, both the above models ignore the underlying temporal dynamics of fashion and is therefore unable to track fashion evolution and understand users' purchasing choice. Reference [10] presented a sparse hierarchical embedding method that utilized visual information for personalized ranking tasks. Reference [21] proposed an One-Class Collaborative Filtering model that jointly models visual appearances and temporal dynamics for estimating users personalized rankings. In their model, item features are divided into visual and non-visual features. Correspondingly, user features are divided into two parts. Both visual and the corresponding user features are time-aware. Since a large corpora of products is involved, modeling in this way will largely increase the quantity of variables. Moreover, it also reduces the consistency of both user and item features and aggravate the existing sparsity issue. Reference [9] is another work using both visual information and temporal dynamics for sequential recommendation. In their work, the authors build a large-scale recommender system to model the dynamics of a vibrant digital art community.

Although the above works have proved that visual information should be used in recommender systems, none of them are designed for rating prediction task. Intuitively, visual appearances help users make purchase decisions rather than influence their ratings. People often need to view the pictures before making a purchase, but give ratings without seeing the pictures again after the products are recieved (except for the case that the actual product's appearances differ a lot from its pictures). It seems unsuited to consider visual appearances for building rating prediction models. However, we see them as an important part of an item's features that affects the users' rating on it. Modeling them makes mining users' common preferences on product appearances more easily and precisely. In addition, those works chose to omit another useful information—review text. As we depict above, review text should be modeled since it can help a RS understand whether a user is attracted by the appearance of an item, what the user mainly cares about, and which aspect of the item contributes most to its ratings. Thanks to this, modeling visual appearances is no longer inappropriate for rating prediction task.

The existing RSs studied either the temporal dynamics of ratings, combination of ratings and reviews, or visual dynamics. To the best of our knowledge, there is no RS model that jointly mines all the three of them. Our work makes use of review text and visual appearances as well as time information to find the reasons behind rating and review evolution, which enables us to capture the temporal drifts of users' preferences and fashion trends.

The rest of the paper is organized as follows. We present background preliminaries in Section II. Our visually-aware temporal rating model with topics is developed in Section III. Section IV presents model fitting. Evaluation results are presented in Section V. The paper is concluded in Section VI.

II. PRELIMINARIES

We begin by briefly describing our previous work named ETDR that jointly mines the temporal changes in user and item latent features together with the associated review text. [15].

The predicted rating $\hat{r}_{u,v}(t)$ for a user u and item v in time period t is calculated according to:

$$\hat{r}_{u,v}(t) = \mu + \beta_u + \beta_v + p_u^T \cdot q_v(t), \tag{1}$$

where μ is an offset parameter, β_u and β_v are user and item biases, and p_u and $q_v(t)$ are K-dimensional user and item factors, respectively. To capture the temporal dynamics in item latent features, the model splits item factors $q_v(t)$ as follows:

$$q_v(t) = q_v + q_{v,t} \tag{2}$$

where q_v is the persistent part of item v's feature and $q_{v,t}$ is associated with each time period. When the model is being fitted, $q_{v,t}$ automatically captures the temporal dynamics of the item's feature in each time period.

To model the review text, q_v and $q_{v,t}$ are linked with topic distributions θ_v and $\vartheta_{v,t}$, respectively.

$$\theta_{v,k} = \frac{e^{\delta \cdot q_{v,k}}}{\sum_{k'=1}^{K} e^{\delta \cdot q_{v,k'}}}, \tag{3}$$

where δ is a parameter which controls the 'peakness' of the transformation and is fit during the learning stage. Thus, for a word w in item v's reviews, it can be sampled as either following θ_v or $\vartheta_{v,t}$ depending on the current word scope distribution $\pi_{v,w}$ and the results of $\sum_{k=1}^{K} \theta_{v,k}\phi_{k,w}$ and $\sum_{k=1}^{K} \vartheta_{v,k}\phi_{k,w}$. Here, $\phi_{k,w}$ is word w's distribution for topic k.

In ETDR model, a document $d_{v,t}$ is defined as the set of the reviews of a particular item v in time period t. Thus, the loss function of the model is a combination of the Mean Squared Error (MSE) of the rating predictions and the (log) likelihood of the review corpus,

$$\begin{aligned} &\mathcal{L}(\mathcal{R}, \mathcal{D}|\Upsilon, \Theta, \Phi, \Delta, z, s) \\ &\quad = \sum_{r_{u,v,t} \in \mathcal{R}} (r_{u,v,t} - \hat{r}_{u,v}(t))^2 - \eta \log p(\mathcal{D}|\Theta, \Phi, z, s), \end{aligned} \tag{4}$$

where z and s are the sets of topic and scope assignments for each word in the corpus $\mathcal{D}$. The detailed definitions of the variables can be found in [15].

III. VISUALLY-AWARE TEMPORAL RATING MODEL WITH TOPICS

We are interested in learning visual temporal dynamics as well as items' non-visual features from ratings and reviews. We believe that visual dynamics play an important role in helping users make purchase decisions. Studying them makes us understand the composition of users' preferences (e.g. who they are and what they like) and their purchase choices.

By accounting for evolving fashion dynamics and review text, we hope to build a model to predict users' preferences on the items in the form of ratings. Formally, we represent the set of users and items with $\mathcal{U}$ and $\mathcal{V}$ respectively. To capture the temporal dynamics, the dataset is divided into $\mathcal{T}$ time periods. $r_{u,v}(t)$ is the rating of user $u \in \mathcal{U}$ on item $v \in \mathcal{V}$ in time period $t \in \mathcal{T}$. $\mathcal{R}$ is the set of all the ratings. Thus, our objective is to predict ratings $\hat{r}_{u,v}(t)$ for user $u \in \mathcal{U}$ on item $v \in \mathcal{V}$ in time period t. The challenge here is to develop efficient methods to make use of the visual information as well as review text to learn fashion trends that are temporally evolving and prediction of users' preferences. The notations used in this paper are listed in Table 1.

A. MODELING VISUAL APPEARANCES

Although modeling review text can improve the accuracy of learned user and item features and give better rating prediction results, it ignores users' preferences toward visual appearances. This factor is especially important when dealing with datasets where visual decision factors are at play.

To model visual appearances of items and exploit users' preferences towards different visual styles, we need to incorporate visual information into the formulation. Previous visual-aware methods handle non-visual features and visual ones separately [21]. In such models, the rating or score for a user-item pair includes the biases, the inner product of the user and item's non-visual features as well as the inner product of their visual features. In those works, the idea is to discover user and item latent features and the visual features separately. However, since new variables are introduced not only for item visual features but also for the corresponding user features, more variables are needed as the number of users grows. Moreover, the type of information contained in items' pictures differs depending on their resolution, shooting angle and other factors. It is inappropriate to separate a user's preferences into two different parts. Upon this understanding, we decompose item v's features q_v into non-visual features and visual features and keep the user features p_u as a whole. Thus, our rating function is remained as Equation (1). We only modify Equation (2) for item feature vector $q_v(t)$ to make it embrace both visual and non-visual information.

TABLE 1. List of notations.

Symbol	Description
$\mathcal{U}$	the set of users in the dataset
$\mathcal{V}$	the set of items in the dataset
$\mathcal{T}$	the set of time periods
$\mathcal{R}$	The set of all ratings in the dataset
$\mathcal{D}$	The set of documents/reviews
$r_{u,v}(t)$	the rating for user u and item v in time period t
$\hat{r}_{u,v}(t)$	predicted rating for user u and item v in time period t
μ	offset parameter
β_u	user bias
β_v	item bias
β_t	bias of time period t
g_v	the Deep CNN features of item v
G	dimensions of g_v
$\mathbf{M}(t)$	embedded matrix captures the temporal dynamics of visual factors in time period t
p_u	K-dimensional user factors
$q_v(t)$	K-dimensional item factors in time period t
q'_v	the *non-visible* features of item v
$q^*_v(t)$	the *visible* features of item v in time period t, which is calculated with Equation (8)
$\boldsymbol{\omega}_v$	a weighting factor for item v to control the extent of its visual feature
θ'_v	*non-visible* topic distribution of item v's review text
$\theta^*_v(t)$	*visible* topic distribution of item v's review text in time period t
$\rho_{v,w,0}$	the probability that word w is chosen as an *non-visible* word
$\rho_{v,w,1}$	the probability that word w is chosen as an *visible* word
$y_{d,i}$	the visibility assignment of the ith word in document d, where $y \in \{0 : \textit{non-visible}\ \text{word}, 2 : \textit{visible}\ \text{word}\}$
$\phi_{z,w}$	Word w's distribution for topic z
α	the hyper-parameters chosen as Dirichlet priors of ρ
σ	the hyper-parameters chosen as Dirichlet priors of ϕ
η	the hyper-parameter that trades off the importance of matix factorization part and the log likelihood of the reviews.
Υ	$\{\mu, \beta_u, \beta_v, \beta(t), p_u, q'_v, \mathbf{M}(t), \boldsymbol{\omega}_v\}$
Θ	$\{\theta'_v, \theta^*_v(t)\}$
Φ	$\{\phi, \rho\}$
Δ	$\{\delta_0 : q'_v \leftrightarrow \theta'_v, \delta_1 : q^*_v(t) \leftrightarrow \theta^*_v(t)\}$
$n^{(-i)}_{k,w'}$	the number of times that $w_{d,i}$ is sampled as with visibility j in all the reviews of item v, excluding the current word assignment $y_{d,i}$

The visual features used in our model are extracted from raw product images with the famous Deep Convolutional Neural Network (i.e., Deep CNN). Let g_v denotes the Deep CNN features of item v, and G represents its dimensions.

We linearly embed the high-dimensional feature vector g_v into the same space as the non-visual features by introducing a $K \times G$ embedding matrix $\mathbf{M}$. Thus, the visual feature of item v is,

$$q_v^* = \mathbf{M} \cdot g_v \tag{5}$$

and item feature q_v becomes,

$$q_v = q_v' + \mathbf{M} \cdot g_v \tag{6}$$

where q_v' is the non-visual feature of v. By learning $\mathbf{M}$ from the data, we can uncover the most useful dimensions of the Deep CNN features for predicting users' preferences and weight them automatically.

Although our model is designed for the datasets where visual decision factors are at play, the contribution of visual factors for each item's feature may be different. Consequently, we add a weighting factor $\boldsymbol{\omega}_v$ for each item to control the extent of its visual feature. q_v is then calculated as

$$q_v = q_v' + \boldsymbol{\omega}_v \cdot \mathbf{M} \cdot g_v \tag{7}$$

Here, the weight factor $\boldsymbol{\omega}_v$, item non-visual feature q_v' and the embedding matrix $\mathbf{M}$ are fit during the learning stage.

B. MODELING VISUAL EVOLUTION AND TEMPORAL DYNAMICS

To uncover the evolving influence of fashion trends on users' preferences and ratings, we are interested in capturing both visual and non-visual dynamics. Considering the sparsity of the datasets, we avoid involving time-dependent parameters for each user or item when developing our model.

1) VISUAL TEMPORAL DYNAMICS

We separate data into $\mathcal{T}$ time periods. To capture the visual dynamics in each time period, we extend our embedded matrix $\mathbf{M}$ as time-dependent. Recall that learning $\mathbf{M}$ can help the model uncover the most useful dimensions of the Deep CNN features for predicting their weights and users' preferences. Since the fashion is changing over time, the weight of each dimension of the Deep CNN features evolves. It is necessary to re-evaluate $\mathbf{M}$ for each time period. So item v's visual feature and whole feature are calculated as

$$q_v^*(t) = \mathbf{M}(t) \cdot g_v \tag{8}$$

and

$$q_v(t) = q_v' + \boldsymbol{\omega}_v \cdot \mathbf{M}(t) \cdot g_v \tag{9}$$

where $\mathbf{M}(t)$ captures the temporal dynamics of visual factors in time period $t \in \mathcal{T}$, and $q_v(t)$ is the item feature of v in the same time period.

2) NON-VISUAL TEMPORAL DYNAMICS

Besides of visual temporal dynamics, there are other time-dependent factors that may affect users' rating behaviors in different time periods. Here, we simply use a time drifting bias to absorb these fluctuates of the ratings. Accordingly, the rating prediction function of our model is

$$\hat{r}_{u,v}(t) = \mu + \beta_u + \beta_v + \beta(t) + p_u^T \cdot q_v(t), \tag{10}$$

where $\hat{r}_{u,v}(t)$ is the predicted rating for user u and item v in time period t, and $\beta(t)$ is the time-dependent bias captures non-visual temporal dynamics in the time period.

C. JOINTLY MODELING FASHION TRENDS AND REVIEW TEXT

Reviews generated by users contain rich information about their preferences and their purchased items' features. These features include both non-visual features and visual ones. To exploit these features will help a RS understand what a user mainly cares about, and which aspect of an item contributes most to its ratings.

To serve this purpose, we develop our model by jointly modeling visual appearances and review text. The assumption is that if an aspect of an item's visual feature matches the fashion trends in a time period, words depicting this aspect will appear more often in the item's reviews in that period; when the fashion changes, the frequency of these words drops. For example, when 'clean classic sneakers' becomes a fashion trend in a time period, a shoe of this style will attract additional customers who keep up with the trends. For those customers, fashion is a high priority, which makes them more willing to share opinions on that topic in their reviews.

To capture these kind of dynamics in reviews, we define document $d_{v,t}$ as the set of all reviews of item v in time period t. We transform item v's visual feature vector $q_v^*(t)$ into *visual* topic distribution $\theta_v^*(t)$. For those non-visual features which are relatively unchanging, we introduce *non-visual* topic distribution θ_v' which is transformed from q_v' to sample the corresponding words. Both transformations are implemented with Equation (3). For each item v and word w in the dictionary, we draw the **word visibility distribution** $\rho_{v,w}$ from a Dirichlet prior α, where $\rho_{v,w,0}$ is the probability that word w is chosen as *non-visible* word and $\rho_{v,w,1}$ is the probability that word w is chosen as *visible* word. In other words, each word i in document d_v is sampled as *non-visible* word or *visible* word according to $\rho_{v,w_{d,i}}$. Note that we are using the same ρ for item v in all time periods. In our point of view, words depicting an item's fashion elements in one or more time periods will have skewed frequency in these periods, therefore be more likely to be selected as *visible* words whose topic distributions are changing over time. On the other hand, the rest of the words who have balanced frequency in each time period will have more chance to be sampled as *non-visible* words. Accordingly, the generative process for document $d_{v,t}$ in the corpus $\mathcal{D}$ is,

1) For each word in dictionary $\mathcal{W}$, choose $\rho_{v,w} \sim$ *Dirichlet*(α), where *Dirichlet*(α) is a Dirichlet distribution with a symmetric parameter α.
2) Transform $\theta_v^*(t)$ from $q_v^*(t)$.
3) Transform θ_v' from q_v'.
4) Choose $\phi_K \sim$ *Dirichlet*(σ)

5) For word position i in $d_{v,t}$,
   a) Choose a topic $z_{v,t,i,0} \sim Multinomial(\theta'_v)$.
   b) Choose a topic $z_{v,t,i,1} \sim Multinomial(\theta^*_v(t))$.
   c) Choose a word $w_{d,i} \sim Multinomial([\rho_{v,*,0} \circ \phi_{z0}, \rho_{v,*,1} \circ \phi_{z1}])$, where $\rho_{v,*,0}$ or $\rho_{v,*,1}$ is a vector consist of the probabilities that each word in $\mathcal{W}$ is chosen as non-visible words or visible words, respectively; $\circ$ is the Hadamard product. Here the subscripts (v, t, i) of z are dropped for brevity.

The graphical model is shown in Fig. 1.

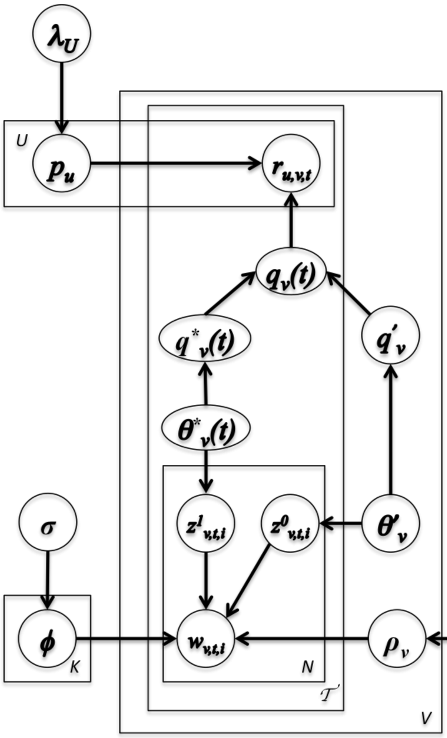

FIGURE 1. Graphical model of VITAL, α and σ are the Dirichlet prior of word scope distribution ρ and word topic distribution ϕ.

For convenience, we define the topic distribution of the set of words with visibility $y \in \{0 : \textit{non-visual}$ word, $1 : \textit{visual}$ word$\}$ in document $d_{v,t}$ as

$$\theta_{v,y} = \begin{cases} \theta'_v, & \text{if } y = 0 \\ \theta^*_v(t), & \text{if } y = 1 \end{cases} \tag{11}$$

Then, the likelihood of the set of all reviews $\mathcal{D}$ is

$$p(\mathcal{D}|\theta', \theta^*, \phi, \rho, y, z) = \prod_{d_{v,t}}^{\mathcal{D}} \prod_{i}^{N_{v,t}} \rho_{v,w_{d,i},y_{d,i}} \cdot \theta_{v,y_{d,i},z_{d,i}} \cdot \phi_{z_{d,i},w_{d,i}} \tag{12}$$

Here $N_{v,t}$ is the number of words in document $d_{v,t}$. Recall that $\phi_{z,w}$ is word w's distribution for topic z. Maximizing the likelihood will make each word's visibility distribution moving towards either θ' or θ^* whose distribution is closer to ϕ_w.

Note that in practice, we do not learn θ'_v and $\theta^*_v(t)$ when fitting the topic model. They are transformed from q'_v and $q^*_v(t)$, respectively. To provide additional flexibility, we define two separate δs, $\Delta = \{\delta_0 : q'_v \leftrightarrow \theta'_v, \delta_1 : q^*_v(t) \leftrightarrow \theta^*_v(t)\}$ as tun-able transformation parameters. However, we find in our experiments that it give better performance to tie them up in the learning stage.

The final model is developed to reflect when ratings are modeled with fashion trends, there should be words in an item's reviews explaining its *non-visual* feature and *visual* feature, respectively. When jointly modeling these two parts, the accuracy of the rating prediction is to be increased. According to this, we define our training objective function as:

$$\mathcal{L}(\mathcal{R}, \mathcal{D}, \mathcal{G}|\Upsilon, \Theta, \Phi, \Delta, y, z) = \sum_{r_{u,v,t} \in \mathcal{R}} (r_{u,v,t} - \hat{r}_{u,v}(t))^2 - \eta \log p(\mathcal{D}|\Theta, \Phi, y, z) \tag{13}$$

where $\mathcal{R}$ is the set of all ratings in the dataset; $\mathcal{D} = \{d_{v,t}|v \in V, t \in \mathcal{T}\}$ is the set of documents/reviews; $\mathcal{G}$ is the set of items' Deep CNN features; $\Upsilon = \{\mu, \beta_u, \beta_v, \beta(t), p_u, q'_v, \mathbf{M}(t), \boldsymbol{\omega}_v\}$ is the set of parameters associated with ratings, including visual embedded matrix in each time period; $\Theta = \{\theta'_v, \theta^*_v(t)\}$ and $\Phi = \{\phi, \rho\}$ are the parameters associated with topics; y and z are the sets of visibility and topic assignments for each word in the corpus $\mathcal{D}$. $\hat{r}_{u,v}(t)$ in the first part of this equation is calculated using Equation (10), and the second part is the log likelihood of the reviews as calculated in Equation (12). η is the hyper-parameter that trades off the importance of the two parts.

IV. MODEL FITTING

Our goal is to simultaneously optimize the rating associated parameters $\Upsilon = \{\mu, \beta_u, \beta_v, \beta(t), p_u, q'_v, \mathbf{M}(t), \boldsymbol{\omega}_v\}$, and topic and visibility associated parameters $\Theta = \{\theta'_v, \theta^*_v(t)\}$ and $\Phi = \{\phi, \rho\}$. We also fit $\Delta = \{\delta_0, \delta_1\}$ at the same time. Therefore, given the set of Deep CNN features $\mathcal{G}$, the ratings $\mathcal{R}$ and reviews $\mathcal{D}$ and their time-stamps, our objective is to find

$$\underset{\Upsilon,\Theta,\Phi,\Delta,z,y}{\textbf{argmin}} \quad \mathcal{L}(\mathcal{R}, \mathcal{D}, \mathcal{G}|\Upsilon, \Theta, \Phi, \Delta, y, z). \tag{14}$$

Since our model contains two separate parts, it is intractable to get the optimal solution directly. Instead, we develop an EM algorithm that alternates between Gibbs sampling [27] and gradient descent, to optimize the parameters. In the E-step, we use Gibbs sampling to learn the word topic distributions and word visibility distributions Φ by fixing the values of Υ and Δ. In the M-step, we perform L-BFGS [28], a quasi-Newton method for non-linear optimization of problems with many variables, to learn Υ and Δ by fixing Φ and each word's $\{y, z\}$ sampled in the last iteration of the E-step.

A. E-STEP

In this step, we update topic distributions Θ using Equation (3) and perform Gibbs sampling to learn the word topic and word visibility parameters Φ by fixing the values of Υ and Δ updated in the M-step. In each iteration, for the ith word in document $d_{v,t}$, we sample $\{z_{d,i}, y_{d,i}\}$ by the probability function of

$$p(y_{d,i}=j, z_{d,i}=k|y_{-i}, z_{-i}, w_{d,i}, \theta', \theta^*(t), \alpha) \propto (n^{(-i)}_{v,w,j}+\alpha_j)\cdot\frac{n^{(-i)}_{k,w}+\sigma_w}{\sum_{w'=1}^{W} n^{(-i)}_{k,w'}+\sigma_{w'}}\cdot\theta_{v,y,k} \qquad (15)$$

where $n^{(-i)}_{v,w,j}$ denotes the number of times that $w_{d,i}$ is sampled as word with visibility j in all the reviews of item v, excluding the current word assignment $y_{d,i}$. We omit the subscript for w in the equation for brevity. $n^{(-i)}_{k,w}$ is the number of times that word $w_{d,i}$ is sampled as topic k in corpus $\mathcal{D}$. Recall that α and σ are hyper-parameters chosen as Dirichlet priors of ρ and ϕ.

Note that for each word in the corpus, the word visibility and topic are sampled simultaneously. The critical difference between our model and previous topic models (such as LDA) is that, instead of sampling topic distribution θ from a Dirichlet distribution, we have multiple topic distributions determined based on Υ for each document, and we use ρ to tune the weights for words of selecting which distribution to follow. Repeated sampling through all the words in the corpus could reach convergence, where the change in ρ and ϕ between successive iterations is sufficiently small. Once the sampling process is finished, we can readily readout the parameters:

$$\rho_{v,w,j} = \frac{n_{v,w,j}+\alpha_j}{\sum_{j'=0}^{1} n_{v,w,j'}+\alpha_{j'}} \qquad (16)$$

and

$$\phi_{k,w} = \frac{n_{k,w}+\sigma_w}{\sum_{w'=1}^{W} n_{k,w'}+\sigma_{w'}}. \qquad (17)$$

Here the notations have the same meanings as in Equation (15), except that the counters n counts all effective samples in the corpus.

B. M-STEP

In this step, we use L-BFGS to learn Υ,Θ and Δ by fixing Φ and $\{y, z\}$ sampled in the last iteration of E-step. Because the variables Θ are linked to Υ, and used in the second part of $\mathcal{L}$, our objective in this step becomes

$$\{\Upsilon, \Theta, \Delta\} = \mathop{\textbf{argmin}}_{\Upsilon,\Theta,\Delta} \quad \mathcal{L}(\mathcal{R}, \mathcal{D}, \mathcal{G}|\Upsilon, \Theta, \Phi, \Delta, y, z), \qquad (18)$$

where Φ, y and z are obtained from the last E-step. Recall that Θ is transformed from Υ and Δ, so they are the actual parameters learned in this step.

We repeatedly perform the EM steps until convergence, i.e., until the changes in Υ and Δ between two consecutive M-steps are sufficiently small, or the maximum number of iterations are performed.

V. EXPERIMENTS

We perform experiments on eleven real-world datasets to investigate the efficacy of our model. We evaluate our model against previous representative rating models, and analyze the influence of hyper-parameters on the rating prediction performance.

A. DATASETS

To evaluate our method on capturing fashion trends and corresponding review text, we are interested in datasets with large corpus of reviews and that visual information plays an important role on users' purchasing choice. We use the public datasets collected from Amazon [29]. The datasets contain $G = 4096$ dimensional visual features extracted from the image of each item. Note that we run our experiments not only on typical visual-determined datasets such as 'Women Clothing' and 'Men Clothing', but also on other datasets like 'Beauty', 'Cell Phone & Accessories' and 'Home & Kitchen' where visual factors have influence on users' preferences according to the commonsense. We keep all the interactions from Feb. 2012 to Jul. 2014 in the datasets, even for those users or items who have only one review. The datasets and their statistics are presented in Table 2.

TABLE 2. Datasets statistics.

Dataset	#users	#items	#reviews
Men Clothing	317,676	70,761	398,048
Men Shoes	323,791	36,529	379,513
Women Clothing	341,157	97,705	460,609
Women Shoes	532,693	81,762	716,945
Beauty	1,006,161	223,486	1,673,549
Cell Phones & Accessories	1,961,960	276,137	2,952,918
Electronics	3,071,616	375,611	5,418,124
Home & Kitchen	1,921,955	342,906	3,156,556
Sports & Outdoors	1,595,127	406,418	2,604,110
Toys & Games	978,032	263,449	1,618,681
Video Games	446,027	38,759	674,561

B. BASELINES

We compare our model with the following state-of-the-art rating models:

- **PMF:** Probabilistic Matrix Factorization (PMF) is a classic latent factor model. It completely ignores the review text and time information. The algorithm can be expressed as Equation (1) but with $q_v(t)$ being time independent. We fit all parameters with L-BFGS.
- ***timeSVD++*:** This is a representative time-aware model [23]. The model divides data into $\mathcal{T}$ time periods, and introduces time-dependent bias and features to capture time-drifting attributes for users and items.
- **HFT:** Hidden Factors and Topics (HFT) is the first model combining ratings and review text for rating prediction task [13]. We set η to 0.1 as suggested in [13].
- **JMARS:** It is another rating prediction model using review text [12]. The model improves the rating prediction performance by capturing the aspect distribution of users' interests in their reviews. The aspect size is set to five in our experiments.

TABLE 3. MSE results for K = 10 of various models and performance improvements of VITAL against other models.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	improvement			
Dataset	PMF	*time*SVD++	JMARS	HFT	TVBPR+	ETDR	VITAL	g vs. a	g vs. b	g vs. c	g vs. d
Men Clothing	1.32353	1.31933	1.32293	1.29991	1.44556	1.29826	**1.29516**	2.14%	1.83%	2.10%	0.37%
Men Shoes	1.26580	1.26697	1.24768	1.24417	1.39193	1.24026	**1.23497**	2.44%	2.53%	1.02%	0.74%
Women Clothing	1.49674	1.50257	1.46937	1.46096	1.67140	1.45983	**1.45935**	2.50%	2.88%	0.68%	0.11%
Women Shoes	1.25202	1.25431	1.22869	1.22166	1.31897	1.22268	**1.22052**	2.52%	2.69%	0.66%	0.09%
Beauty	1.53418	1.53417	1.52420	1.52764	1.65441	1.52902	**1.52194**	0.80%	0.80%	0.15%	0.37%
Cell Phones & Accessories	1.89343	1.89023	1.86643	1.87374	2.08974	1.86737	**1.86501**	1.50%	1.33%	0.08%	0.47%
Electronics	1.67276	1.67189	1.65786	1.66214	1.83668	1.66047	**1.65609**	1.00%	0.95%	0.11%	0.36%
Home & Kitchen	1.50708	1.51263	1.49399	1.49111	1.67560	1.49069	**1.48916**	1.19%	1.55%	0.32%	0.13%
Sports & Outdoors	1.35914	1.35842	1.34785	1.34695	1.52339	1.34498	**1.34374**	1.13%	1.08%	0.30%	0.24%
Toys & Games	1.31383	1.31581	1.30709	1.30487	1.49508	1.29552	**1.29277**	1.60%	1.75%	1.10%	0.93%
Video Games	1.62430	1.63145	1.60477	1.60347	1.80495	1.60645	**1.59935**	1.54%	1.97%	0.34%	0.26%
Average								1.67%	1.76%	0.62%	0.37%

TABLE 4. The standard error of the mean squared errors for K = 10 of the models.

	(a)	(b)	(c)	(d)	(e)	(f)	(g)
Dataset	PMF	*time*SVD++	JMARS	HFT	TVBPR+	ETDR	VITAL
Men Clothing	0.0119	0.0120	0.0118	0.0119	0.0128	0.0116*	0.0117
Men Shoes	0.0120	0.0121	0.0119	0.0118*	0.0130	0.0118*	0.0118*
Women Clothing	0.0107	0.0106	0.0105	0.0105	0.0124	0.0104	0.0102*
Women Shoes	0.0084	0.0084	0.0081*	0.0081	0.0087	0.0082	0.0082
Beauty	0.0063	0.0063	0.0062*	0.0062*	0.0067	0.0062*	0.0062*
Cell Phones & Accessories	0.0033*	0.0034	0.0033*	0.0033*	0.0035	0.0033*	0.0033*
Electronics	0.0020	0.0020	0.0019*	0.0019*	0.0020	0.0019*	0.0019*
Home & Kitchen	0.0032	0.0032	0.0032	0.0031*	0.0035	0.0031*	0.0031*
Sports & Outdoors	0.0041*	0.0042	0.0041*	0.0041*	0.0045	0.0041*	0.0041*
Toys & Games	0.0060	0.0060	0.0058	0.0058	0.0065	0.0057*	0.0057*
Video Games	0.0097	0.0096*	0.0098	0.0096*	0.0108	0.0097	0.0096*

- **TVBPR+:** We also run experiments for the state-of-the-art visual-aware model named TVBPR+ which is proposed in [21]. We use the preference predictor $\hat{x}_{u,i}$ which is introduced in that paper as the rating prediction function and fit all the parameters with L-BFGS. When we transform the TVBPR+ function to fit the rating prediction task, we substitute $\hat{x}_{u,j}$ with a real value rating, and reverse the sign of the regularizer. The goal is to minimize the mean squared errors and learn the parameters. Both latent feature dimension K and the dimension of visual style space K' are set to 10. The target function of the model is converted to the sum of MSE with $L2$ regularizer and λ is set to 0.0001.
- **ETDR:** The model considers ratings and reviews as well as temporal dynamics in the learning stage [15]. The ratings are predicted according to Equation (1). And the loss function can be expressed as Equation (4).

C. TRAINING

We randomly divide each dataset into training, validation, and test sets. We use 80% of each dataset for training, and evenly split the rest into validation set and test set. The latent factors and topic of each word are initialized uniformly at random. The visibility distribution are initialized according to the hyper-parameter α. In our experiments, we only report the results when α is set to $\{3, 3\}$. We set time period length to be 30 days for all the datasets. K is set to 10 for all models. We have tuned the hyperparameters for each model to get the best results. Although our model can achieve the best results within 1000 iterations on all the tested datasets, we run for more than 3000 iterations for other models on each dataset to ensure they get the best results. After every 10 iterations of gradient descent, we run 10 iterations of Gibbs sampling to optimize Φ and compute Mean Squared Error (MSE) on the validation and test sets. We report the MSE on the test set for each model with parameter setting that achieves the lowest error on the validation set.

D. PERFORMANCE

Results in terms of MSE are shown in Table 3. The lowest MSE on each dataset are marked with bold font. The right side shows the improvement of VITAL against other models.

As is shown in table 3, timeSVD++ performs better than PMF on several datasets such as 'Men Clothing' and 'Cell Phones & Accessories', but the average is worse. This indicates that only considering temporal dynamics is not sufficient, even for the datasets with obvious time-dependent factors like fashion. In most cases, rating models with topics can get lower MSE than the others. This shows the importance of considering review text when predicting ratings. When compared with other baseline models, our model achieves the best performance on all the tested datasets. This is due to the capability of capturing visual dynamics which make our model be able to recognize the fashion elements in each time period.

To illustrate the stability of the ratings predicted by the models, we give the standard error of the MSEs in Table 4. From the table, we can see that VITAL achieve the smallest standard errors on almost all the datasets except for Men Clothing and Women Shoes. This shows that VITAL performs stabler on predicting ratings than the other methods.

In the experiment, we find that TVBPR+ does not perform well in predicting ratings. The reason is that it is designed for the objective of maximizing the differences between positive and negative samples, but not minimizing the mean squared errors between the predicted scores and the real value ratings. Since they have totally different goals, the score functions are designed in different ways. Therefore, it is not applicable to directly use it to predict ratings. It is necessary to develop a new visual-aware model for rating prediction tasks.

E. ANALYSIS

Besides the MSE performance, we are also interested in the hyper-parameter and maximum iteration settings of our model. We believe that correct settings will help the model get better predictions in less time.

TABLE 5. Relationship between #ratings and #iterations performed to get optimal solution.

Datasets	#Ratings	#Users + #Items	Iterations
Men Shoes	379,513	360,320	210
Men Clothing	398,048	388,437	330
Women Clothing	460,609	438,862	200
Video Games	674,561	484,786	960
Women Shoes	716,945	614,455	210
Toys & Games	1,618,681	1,241,481	1000
Beauty	1,673,549	1,229,647	580
Sports & Outdoors	2,604,110	2,001,545	950
Cell Phones & Accessories	2,952,918	2,238,097	990
Home & Kitchen	3,156,556	2,264,861	980
Electronics	5,418,124	3,447,227	970

1) #RATINGS vs. #ITERATIONS

Table 5 shows the relationship between *#ratings* and *#iterations* performed to get optimal solution. The numbers of iterations are shown in the rightmost column. We also give the summation of *#users* and *#items* in each dataset. From the table we can see there is a strong relationship between these two numbers. As the number of ratings grows, more iterations has to be performed. This is for the reason that datasets containing more ratings are often accompanied by larger number of users and items who generate these ratings. And the summation of *#users* and *#items* is proportional to the number of parameters learned in the iterative process. Additionally, more ratings means there are larger number of documents which also need more times to run gibbs sampling process.

We notice that there are two exceptions in the table—Video Games and Beauty. The former has only 674,561 ratings whereas 960 iterations are performed to reach optimal. For the latter we only run 580 iterations for 1,673,549 ratings. By analyzing the intermediate results generated in the iterative process, we found although the energy (value) of the function decreased continuously, the MSEs on the validation sets were fluctuating. This may cause uncertainty in some degree like what happened on the datasets of Video Games and Beauty, but will not affect the overall trend.

2) HYPER-PARAMETER η

It is an important parameter in our model that trades off the impact of the rating and topic parts. Although in the experiments we fixed the parameter to 0.5 for comparison, it is necessary to know how to choose it for different datasets.

Fig. 2 shows the MSE results of using different η for the datasets of 'Men Shoes', 'Women Shoes' and 'Women Clothing'. As can be seen from the figure, the MSEs of both Shoes datasets are not smooth when η is smaller than 0.4, but decrease to a steady level after 0.5. This indicates that when running VITAL for these datasets, a large η might be needed (at least larger than 0.4). In contrast, the impact of varying η on the MSE of Women Clothing is very small. This implies distinct η for each dataset should be given for better prediction results in practice.

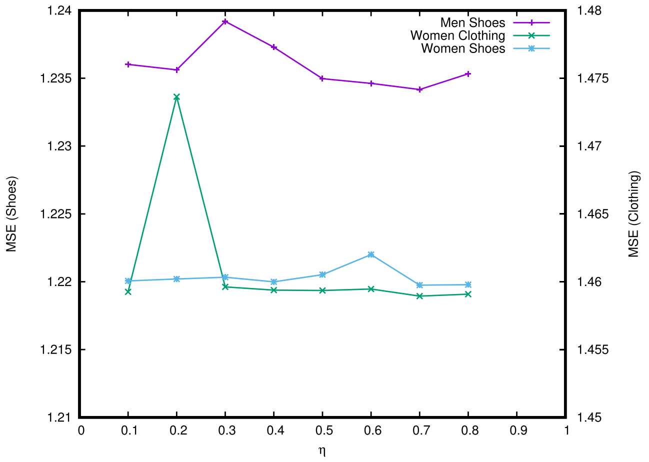

FIGURE 2. MSE results when varying hyper-parameter η on datasets of Shoes and Women Clothing.

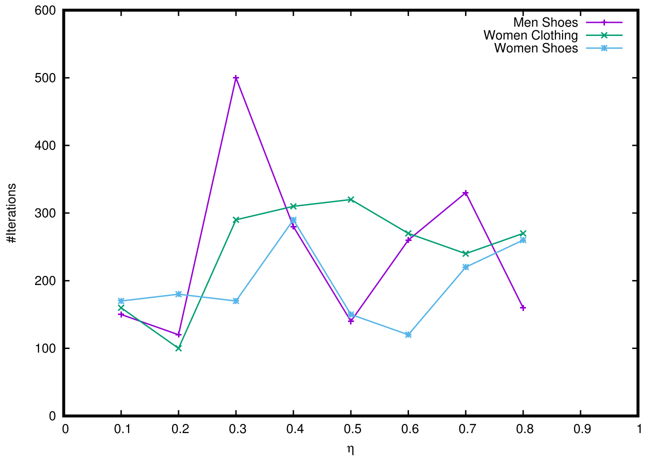

FIGURE 3. The number of iterations to obtain the best MSE results on validation sets at various η.

Fig. 3 gives the numbers of iterations performed for those MSE results shown in Fig. 2. Recall that from Table 5 we find as the number of ratings grows, more iterations are performed. However, according to the figure, no general trend can be found to explain the fluctuations of varying η for each dataset. It seems completely random. In spite of that, the numbers of iterations of these small datasets (#ratings less than 1,000,000) are still within a certain range, not exceeding 500. It seems that the number of the ratings is a more important factor than the value of η when setting maximum iteration times for a dataset.

VI. CONCLUSION

In this paper, we present VIsually-aware Temporal rAting modeL with topics (VITAL) that mines user and item latent features from both item visual appearances and reviews. Our model jointly mines the temporal changes in items' visual factors and their latent non-visual features together with the associated review text in a single learning stage. We run experiments on eleven real-world public datasets and the results show that our model outperforms the state-of-the-art RSs, such as PMF, *time*SVD++, HFT, JMARS, ETDR, and TVBPR+ on rating prediction accuracy. Through hyper-parameter analyzing, we demonstrate η plays an important role in balancing rating and topic parts and should be chosen for each dataset distinctively.

REFERENCES

[1] M. Gorgoglione, U. Panniello, and A. Tuzhilin, "The effect of context-aware recommendations on customer purchasing behavior and trust," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 85–92.

[2] A. Krohn-Grimberghe, L. Drumond, C. Freudenthaler, and L. Schmidt-Thieme, "Multi-relational matrix factorization using Bayesian personalized ranking for social network data," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, New York, NY, USA, 2012, pp. 173–182.

[3] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[4] B. Marlin, R. S. Zemel, S. Roweis, and M. Slaney. (Jun. 2012). "Collaborative filtering and the missing at random assumption." [Online]. Available: https://arxiv.org/abs/1206.5267

[5] J. J. McAuley and J. Leskovec, "From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 897–908.

[6] L. Xiang *et al.*, "Temporal recommendation on graphs via long-and short-term preference fusion," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 723–732.

[7] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. AAAI*, 2016, pp. 144–150.

[8] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, and N. Sundaresan, "Large scale visual recommendations from street fashion images," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1925–1934.

[9] R. He, C. Fang, Z. Wang, and J. McAuley. (Jul. 2016). "Vista: A visually, socially, and temporally-aware model for artistic recommendation." [Online]. Available: https://arxiv.org/abs/1607.04373

[10] R. He, C. Lin, J. Wang, and J. McAuley. (Apr. 2016). "Sherlock: Sparse hierarchical embeddings for visually-aware one-class collaborative filtering." [Online]. Available: https://arxiv.org/abs/1604.05813

[11] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4642–4650.

[12] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang, "Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS)," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 193–202.

[13] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 165–172.

[14] G. Ling, M. R. Lyu, and I. King, "Ratings meet reviews, a combined approach to recommend," in *Proc. 8th ACM Conf. Recommender Syst.*, 2014, pp. 105–112.

[15] Y. Liu, Y. Liu, Y. Shen, and K. Li, "Recommendation in a changing world: Exploiting temporal dynamics in ratings and reviews," *ACM Trans. Web*, vol. 12, Feb. 2017, Art. no. 3.

[16] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 171–180.

[17] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 618–626.

[18] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 56–65.

[19] K. Bauman, B. Liu, and A. Tuzhilin, "Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 717–725.

[20] G. Xun, Y. Li, J. Gao, and A. Zhang, "Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 535–543.

[21] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 507–517.

[22] P. B. Kantor, L. Rokach, F. Ricci, and B. Shapira, *Recommender Systems Handbook*. New York, NY, USA: Springer, 2011.

[23] Y. Koren, "Collaborative filtering with temporal dynamics," *Commun. ACM*, vol. 53, no. 4, pp. 89–97, 2010.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[25] X. He, T. Chen, M.-Y. Kan, and X. Chen, "TriRank: Review-aware explainable recommendation by modeling aspects," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1661–1670.

[26] G.-N. Hu, X.-Y. Dai, Y. Song, S.-J. Huang, and J.-J. Chen. (Jan. 2016). "A synthetic approach for recommendation: Combining ratings, social relations, and reviews." [Online]. Available: https://arxiv.org/abs/1601.02327

[27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.

[28] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Math. Comput.*, vol. 35, no. 151, pp. 773–782, 1980.

[29] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 43–52.

YINING LIU received the M.S. degree in computer science from the Dalian University of Technology, Liaoning, China, in 2011, where he is currently pursuing the Ph.D. degree. He was a Visiting Ph.D. student with the NYU Tandon School of Engineering, from 2013 to 2015. His research interests include recommender systems and collaborative filtering algorithms.

YANMING SHEN received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2000, and the Ph.D. degree from the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, in 2007. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology, China. His general research interests include packet switch design, data center networks, peer-to-peer video streaming, and algorithm design, analysis, and optimization. Prof. Shen was a recipient of the 2011 Best Paper Award for Multimedia Communications awarded by the IEEE Communications Society.