ATLANTIS PRESS

*Atlantis Studies in Uncertainty Modelling, volume 1*

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

# Modeling Implicit Feedback and Latent Visual Features for Machine-Learning Based Recommendation

**Yue Guan, Qiang Wei[1], Guoqing Chen, Xunhua Guo**

China Retail Research Center, School of Economics and Management, Tsinghua University, Beijing, 100084

## Abstract

To leverage the rapid accumulation of rich media on the Internet, this paper proposes a Multi-View Bayesian Personalized Ranking (MVBPR) recommendation model, which combines visual and textual content, along with uncertainty modeling of consumer preference in form of implicit feedback and visual representation in form of latent factors. MVBPR is a machine-leaning framework integral of deep-learning (i.e., SCAE) and topic modeling (i.e., LDA) strategies to fuse image and text information. Moreover, extensive experiments demonstrate MVBPR's advantages over baseline models, including its superiority in dealing with the cold start situation.

**Keywords:** Multi-view information, Machine learning, Stacked convolutional auto-encoder, Topic modeling, Information fusion

## 1 Introduction

With the prosperity of information technology and mobile Internet in recent years, a large number of social and economic activities rely heavily on the support of information. Many kinds of information systems and services arise, from industry-level infrastructure, e-commerce website to community and personalized applications. According to the statistics (https://www.domo.com/blog/data-never-sleeps-4-0/), in every minute, Amazon complets $222,283 in sales, YouTube users share 400 hours of new video, Instagram users like 2,430,555 posts, Siri answer 99,206 requests, etc. On one hand, the abundant information enriches our lives and provides us with more freedom to choose the content we like. On the other hand, the information overload problem naturally emerges and generates frustration and confusion for decision making [2].

To alleviate the problem, recommender systems are widely applied in practice and play an important role in supporting users' decision processes. For instance, with historical data, recommendation systems [1] can help infer user preferences and accordingly provide accurate recommendation services, which may reduce consumers' decision cost, shorten the decision time and improve their satisfaction.

The development of deep learning techniques (e.g., convolutional neural networks) makes large-scale unstructured and visual content analysis a trendy and viable research topic [9][11][22]. This injects great potential for recommender system design. Despite the large amount of work in recommender system design [1], previous studies seldom consider utilizing the abundant visual data or combine multiple forms of data together to achieve better performance. To incorporate the contextual information, there are two kinds of uncertainties that need to be addressed and has not been sufficiently considered. We deem these issues important for improving consumer shopping experience as they are highly relevant with consumer shopping behavior.

First, there is uncertainty in consumer preferences, because preferences are usually not explicitly stated and can only be implicitly inferred from online behaviors. For instance, if a consumer clicks or purchases a product, it is most likely that the consumer prefers the product to the unclicked/unpurchased ones. This is the underlying assumption of our work, while the preference order within the group of purchased products or unpurchased products remains unknown. Thus, a pairwise learning model can be trained to automatically learn the products' ranking for each consumer to deal with the preference uncertainty.

Second, there is uncertainty in the perception effects of product images. For online consumers, images are crucial information as they could affect consumers' first impressions of a product [14]. Compared with texts, images are deemed more convincible and more expressive with enriched and detailed information [20][24]. Meanwhile, heterogeneity in consumers also exists, reflecting that even confronted with the same image, different people may have different perspectives and perceptions.

From a technical point of view, an image is composed of numerical pixel values without semantic meanings. Existing efforts on image processing mainly focus on tasks such as image classification, segmentation, and object detection [12]. Nevertheless, prior research seldom investigates the subtleties within a specific product category, which, however, may significantly impact consumers' online shopping behaviors. Therefore, how to represent and extract the information from images properly and incorporate it

[1] Corresponding author: weiq@sem.tsinghua.edu.cn

into the recommendation model effectively is one of the key challenges in this study.

In addition to images, consumers often tend to embrace different sources of information before making a purchase, e.g., consumers may want to know opinions from other buyers, especially when shopping for experience goods. Prior research has shown that product reviews could significantly impact sales [4][28]. Moreover, this kind of User Generated Content (UGC) could be more convincing compared with Marketer Generated Content (MGC) [13].

Built on above-mentioned points, a machine-learning based model, namely, Multi-View Bayesian Personalized Ranking recommendation model (MVBPR), is proposed which uses stacked convolutional auto-encoder and topic modeling techniques to extract information from both visual and textual data. Then implicit feedbacks and latent features are distilled and integrated into a ranking-based recommendation framework. The rest of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the model framework and formulation. Section 4 demonstrates the outperformance of MVBPR with extensive data experiments. Section 5 concludes the paper.

## 2 Literature Review

### 2.1 Recommender systems

Matrix factorization [10] has been widely applied in Netflix movie recommendations. Its basic idea is to decompose the rating matrix into two small matrices, namely, item latent matrix and user latent matrix. The missing data in the rating matrix could be inferred through the interaction of the two matrices. Though easy and effective in the movie recommendation context, it easily becomes overfitting when the rating matrix is sparse.

Rating, as a typical explicit feedback, is usually sparse in reality, while implicit feedback data is more accessible for recommender systems nowadays. Bayesian personalized ranking [21] is the state-of-the-art recommendation model for implicit feedback. It directly optimizes the ranking of items using BPR-OPT. Combined with matrix factorization, BPRMF outperforms a variety of competitive baselines, which also acts as one of the baseline models in our study. Another baseline algorithm MMMF uses hinge loss function and has similar performance to BPRMF as proved by [8].

To further boost recommendation performance, hybrid algorithms that combine matrix factorization and context information are proposed. Liu et al. [15] analyze review content to elicit user preferences on different aspects of a product. Some studies propose the Latent Dirichlet Allocation to model unstructured text and enhance the performance of article or product recommendation [19][23]. Other hybrid methods utilize auxiliary information, such as user demographics, social relations, social network, and product description [6][15][16][17] to improve recommendation performance to a further extent. Nevertheless, these studies did not consider rich media content, which is one of the focus in this study.

### 2.2 Uncertainty modeling of visual data

As an essential part of deep learning technology, convolutional neural networks have gained great popularity for image processing related tasks and achieved superior performance compared with traditional algorithms [12]. They eliminate the tedious and ineffective feature engineering work, treat raw data as input, and automatically learn high-level features from low-level input. Their special network structures could preserve the input's neighborhood relations and spatial locality in their latent higher-level feature representations, which largely contributes to their remarkable performance [8][9][11].

Several deep learning networks are proposed to extract features from images. Stacked auto-encoder is an unsupervised model that consists of encoder and decoder, aiming to obtain a low dimensional representation, and at the same time reconstruct the original input [18]. Wang et al. [26] propose a generalized Bayesian stacked denoising auto-encoder (SDAE) to extract features from unstructured text and use this as prior information to guide the learning of latent features, and then generate recommendations based on probabilistic matrix factorization. Some research efforts use the pre-trained Alexnet [11] or VGG-16 [22] models to generate features with the last classification layer removed. Through matrix embedding operation upon the output, an effective feature representation can be obtained specific to the recommendation context [8][25]. For instance, He and McAuley [8] utilize the output of trained Alexnet model, which is a 4096-dimension feature. This high dimensional feature is converted to a low dimensional latent feature with embedding approach. However, as different contexts may need different feature learning models, a trained model in image classification task cannot ensure effective feature extraction in the product recommendation context.

### 2.3 Information fusion

Another related stream of literature is by Zhang et al. [27], in which they propose a collaborative knowledge embedding model that leverages image, text and structural relationship in a single Bayesian probabilistic matrix factorization model, to recommend movies/books. Visual embedding and textual embedding are implemented through two auto-encoder structures, namely, stacked convolutional auto-encoder (SCAE) and stacked denoised auto-encoder (SDAE). To integrate multiple sources of information more effectively, Guan et al. [7] make a further step and design an integration framework through the combination of auto-encoders and

embedding approaches. These two models are trained end-to-end and the extracted features lack enough understandability as per the "black box" criticism of deep neural networks.

# 3 Model Framework

## 3.1 Problem Formulation

As mentioned in Section 1, to reconcile the uncertainties involving consumer implicit preference and image perception, this paper proposes a Multi-View Bayesian Personalized Ranking model framework that mainly includes two parts: content extraction and information fusion, to integrate visual, textual, and implicit feedback content simultaneously, thus offer a personalized recommendation solution.

Specifically, in online shopping context, the implicit feedback matrix is denoted as $R$. If consumer $u$ has bought product $i$, then $R_{ui} = 1$, otherwise 0. Given the multi-view information context (i.e. images, descriptions and reviews) and the implicit matrix $R$, the research problem is to generate a recommendation list for each consumer on those products that have not received any feedback from this focal consumer.

The model framework is as shown in Figure 1 and will be detailed in Subsections 3.2 and 3.3.

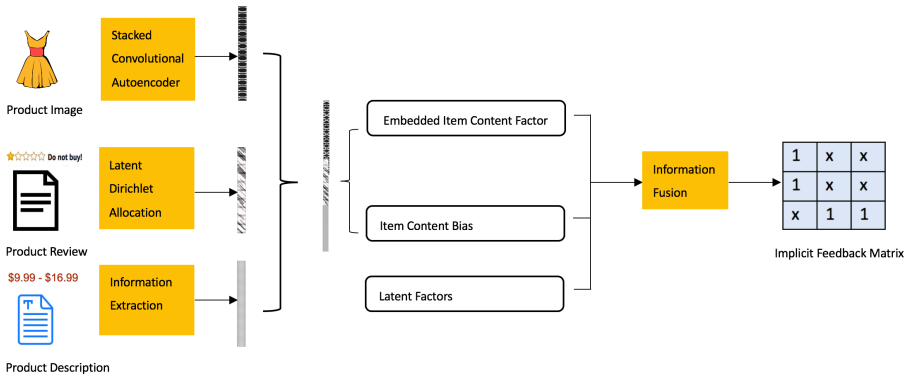


Figure 1: MVBPR model

## 3.2 Content information extraction

The content information in the model comes from three sources: images, descriptions and reviews. A 6-layer stacked convolutional auto-encoder is designed to extract visual features from images. Topic modeling approach is deployed to obtain the topic distribution of the review content. As descriptions are in structured textual form, we conduct simple text processing to acquire the product properties.

### 3.2.1 Visual features

Auto-encoder is a kind of unsupervised neural network that can generate a low dimensional latent feature representation for high dimensional input. It aims to reconstruct the image through the encoder and decoder parts. The output of the encoder keeps discriminative information and discards irrelevant noises, which is consistent with the objective of our uncertainty modeling context.

In spirit of the work in Masci et al. [18], this paper proposes a 6-layer stacked convolutional auto-encoder (SCAE). The first three layers constitute the encoder and the last three layers constitute the decoder (Figure 2). This is a symmetric structure with the encoder and decoder sharing the same weight matrices. Layers 3 and 4 are fully connected layers and the other layers are convolutional and deconvolutional layers. After convolution operation in layer 1, image size is decreased to ¼ (stride = 2), and the number of feature maps increases from 3 to 64. The second convolutional layer keeps the same feature map number (stride = 1) and the third fully connected layer has 128 nodes connecting the encoder and decoder. Deconvolutional layer 4 to layer 6 do the inverse operations. After layer 6, image size returns to the original size with feature map number equal to 3.

Convolutional operation can preserve the input's neighborhood relations and spatial locality in their latent higher-level feature representations [12]. Furthermore, auto-encoder structure provides an efficient and effective image representation with a low dimensional feature. Therefore, this funnel-like convolutional auto-encoder combines the advantages of the two network structures.

More specifically, in the formulation of SCAE, for a convolutional layer, the input of layer $l+1$ (output of layer $l$) is defined as

$$X_{l+1} = \sigma(W_l * X_l + b_l), \quad (1)$$

where $*$ denotes convolution operation and $\sigma$ denotes the sigmoid function.

For a fully connected layer, the input of layer $l+1$ (output of layer $l$) is defined as

$$X_{l+1} = \sigma(W_l \cdot X_l + b_l) \quad (2)$$

The objective function is to minimize reconstruction error with regularization for the weight matrices $W_i$ as in Eq. (3).

$$arg \min_{W_i, b_i} L = \|X_6 - X_0\|^2 + \lambda \sum_{i=1}^{l/2} \|W_i\|^2 \quad (3)$$

As shown in Figure 2, the output of layer 3, namely, $X_3$, is the visual features that will be utilized in the recommendation model.


Figure 2: 6-layer stacked convolutional auto-encoder

#### 3.2.2 Review features

Topic modeling is a statistical modeling method that allows observational documents to be described by the distribution of hidden topics. Latent Dirichlet Allocation (LDA) is a typical topic model [3].

LDA aligns well with our research context because of the following reasons. Product reviews are unstructured texts in varied lengths generated by consumers with different language styles. Consumers also have heterogeneous product experiences, which makes it difficult to summarize the reviews with an exact depiction. Similarly, LDA is a probability-based algorithm that represents documents (here all the reviews of a product are regarded as a document) with different distributions on a series of topics.

The review generation process [3] is as follows.

(1) Choose topic distribution for product $i$ $\theta_i \sim Dir(\alpha)$
(2) Choose review word distribution for topic $k$ $\varphi_k \sim Dir(\beta)$
(3) For each word $j$ of product $i$'s review,
   (a) Choose the review topic from product-topic distribution $z_{ij} \sim Multinomial(\theta_i)$
   (b) Choose the review word from topic-word distribution $w_{ij} \sim Multinomial(\varphi_{z_{ij}})$

There are some common topics that consumers may mention when posting reviews, such as quality, delivery, and price. At the same time, different products may exhibit different patterns on the topic distribution. For instance, consumers may care more about the price for product A, while focusing the material most for product B. The topic distribution obtained by our topic modeling approach serves as an important UGC product feature.

#### 3.2.3 Description features

On e-commerce platforms, product descriptions are an inevitable part of information together with product images. The content covered in the description usually includes price, available date, material, weight, brand, etc. This kind of structural information serves as the basic information provided by sellers and is also of importance to consumer decision-making. Thus, description information is added into our recommendation framework. Relevant features can be extracted with text processing technique, which will be introduced in Section 4.

### 3.3 Information fusion and recommendation

For a focal product $j$, suppose the visual, review, description features obtained in Section 3.2 are denoted as $v_j^m \in R^m$, $v_j^r \in R^r$, $v_j^d \in R^d$. Note that the features can be of different dimensions. $v_j^c$ is a concatenation of the feature vectors, i.e., Eq. (4).

$$v_j^c = [v_j^m, v_j^r, v_j^d] \tag{4}$$

A matrix embedding operation is imposed on the content feature $v_j^c$, namely, $\overline{v_j^c} = E_c \cdot v_j^c$, where $E_c$ is $K \times C$ embedding matrix that transforms a $C$ dimensional feature into a $K$ dimensional feature ($K < C$). The intuition behind the embedding operation is that there may be some redundant information contained in the three kinds of features [7]. For instance, both reviews and descriptions may contain the size information of a product. The embedded feature can merge multiple sources of information in a more compact and efficient format.

To be consistent with prior literature [8], the preference of consumer $u$ to product $i$ can be formulated as Eq. (5).

$$\widehat{x_{u,i}} = \alpha + \beta_i + \beta_u + \gamma_u^T \gamma_i + \theta_u^T (E_c v_i^c) + \beta'^T v_i^c \tag{5}$$

where $E_c v_i^c$ is the embedded content feature, $\theta_u^T$ represents the consumer's preference over the content features. $\beta'^T v_i^c$ is the content bias term which serves the same role as latent bias term $\beta_i$, $\beta_u$.

According to Bayesian Personalized Ranking (BPR) [21], the overall optimization objective is as Eq. (6).

$$\sum_{(u,i,j) \in D_S} \ln \sigma(\widehat{x_{uij}}) - \lambda_\Theta \|\Theta\|^2 \tag{6}$$

$D_S$ is the training set consisting of triple $(u, i, j)$ representing consumer $u$ prefers product $i$ to product $j$. $\widehat{x_{uij}} = \widehat{x_{ui}} - \widehat{x_{uj}}$.

The overall model is named as Multi-View Bayesian Personalized Ranking (MVBPR). Furthermore, MVBPR degenerates to three single models when only one of the three features is available, denoted as MVBPR-M (images only), MVBPR-R (reviews only) and MVBPR-D (descriptions only). Extensive experiments in Section 4 reveal the outperformance of MVBPR and shed light on how different features contribute to the overall recommendation performance.

## 4 Experiments and Results

### 4.1 Experiment settings

To demonstrate the effectiveness of the proposed model, experiments on real-world dataset (www.amazon.com) were conducted. Specifically, the Women Dress category was chosen. It contained about 20,000 products with corresponding images, descriptions and reviews until May 2017. Women dresses are typical experience goods, where images and reviews are perceived to be critical in decision-making.

Area Under the ROC Curve (AUC) was chosen as the evaluation metric, which is defined as $AUC = \frac{1}{|U|}\sum_u \frac{1}{|P(u)|}\sum_{(i,j)\in P(u)} \delta(\widehat{x_{ui}} > \widehat{x_{uj}})$ , where $P(u) = \{(i,j)|R(u,i) = 1 \text{ and } R(u,j) = 0\}$ , $\delta(a)$ is an indicator function that equals 1 if $a$ is true.

The following baseline models were chosen for comparison purpose.

(1) Most Popular (MP). This model ranks products in the order of their sales performance, which is a non-personalized recommendation algorithm.
(2) BPRMF. A pairwise personalized ranking model proposed by Rendle et al. [21], which utilizes implicit feedback information without considering content information.
(3) MMMF. The difference of MMMF from BPRMF is that MMMF uses a hinge loss rather than sigmoid function in the optimization objective[5].
(4) MVBPR-M/R/D. The three single information source based models degenerate from the proposed MVBPR model, focusing only on images, reviews or descriptions, respectively.

## 4.2 Dataset preparation

Data preprocessing was conducted before the experiments. Consumers who had written less than 5 reviews and products which had no reviews were deleted, after which the dataset was divided into training set, validation set and test set in the following steps. For each product, one positive feedback (i.e., purchased) was randomly selected into the validation set, another one was selected into the test set, and the other positive feedbacks were selected into the training set. Products unpurchased were treated as negative feedback. Finally, the dataset contained 1,893 products, 997 consumers and 6,300 purchase records.

Two test sets were prepared to evaluate the model performance, including a full test set and a cold start test set. The full test set consisted of the whole test set, while the cold start test set consisted of the products that had few reviews, i.e., less than 5 reviews. The reason for that is because the cold-start problem is a critical issue in recommender systems. Since the proposed model takes account of various information, it should possess the ability to alleviate the cold start problem to some extent.

Model hyperparameters were chosen using validation set (Table 1) and performance evaluation was conducted on the test sets.

| Model | Regularization Parameter for General Terms | Regularization Parameter for Embedding Matrix | Regularization Parameter for Bias Terms |
|---|---|---|---|
| MMMF | 100 | --- | 0.1 |
| BPRMF | 100 | --- | 0.01 |
| MVBPR | 10 | 0.001 | 0.001 |
| MVBPR-M | 10 | 0.001 | 0.001 |
| MVBPR-R | 10 | 0.001 | 0.001 |
| MVBPR-D | 10 | 0.001 | 0.001 |

Table 1: Hyperparameter settings

## 4.3 Content information extraction results

### 4.3.1 Visual features

The SCAE proposed in Section 3 was trained with Tensorflow (https://www.tensorflow.org) framework on a GPU server. To validate and visualize the information effectiveness of the latent visual features, we reconstructed the images with the latent features and optimized model parameters and compared them with the original images as in Figure 3. There are three group of images. In each group, the first rows are original images, and the second rows are the reconstructed images by SCAE.

It can be observed that the key discriminative and representative characteristics were captured in the reconstructed images except for some color properties, which confirmed the appropriateness of the visual feature component.

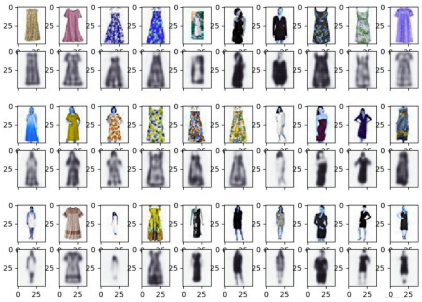

Figure 3: Visualization of SCAE

### 4.3.2 Review features

In the review modeling component, the total topic number was set as 32. Prior Dirichlet distribution hyperparametes were set as $\alpha = 1.6$ , $\beta = 0.01$ . Generally speaking, consumers would not read all the reviews of a product. On mainstream e-commerce platforms, reviews with the most helpful votes are usually displayed at the top of a review page. Therefore, only reviews with at least one helpful vote were selected into the documents.

The LDA model was optimized with Gibbs Sampling [3]. Some randomly selected topics and their representative words (words with the largest probability in a topic) were displayed in Table 2. It could be observed that the words within a topic were semantically similar, while words across topics were dramatically different, indicating that LDA is highly applicable in the context.

| Topic No. | Topic | Top 10 words |
|---|---|---|
| Topic #1 | Size | Small, medium, fit, large, review, short, run, perfectly, super, normally |
| Topic #2 | Delivery | Arrived, time, shipping, expected, seller, fast, price, received, happy, week |
| Topic #3 | Clothes parts | Back, front, seam, side, didn't, big, picture, bottom, star, hole |
| Topic #4 | Experience | Comfortable, soft, comfy, fit, super, fabric, casual, material, flowy, perfectly |

Table 2: Representative review topic samples

#### 4.3.3 Description features

The product properties extracted from descriptions were *Product Price, the First Available Date on Amazon and Material*. Other properties such as *product weight* suffered from data missing problem, so they were not included. *Material* was a categorical variable including 18 classes, and 20 features were obtained in total as in Table 3. All the feature values were standardized to ensure they fall in $[0,1]$.

| Property | Type | Example |
|---|---|---|
| Price | Numerical | $23.99 |
| First avail. date | Date | 8-Oct-2016 |
| Material | Binary | cotton, polyester, spandex, rayon, modal, nylon, viscose, lyocell, linen, chiffon, demin, cashmere, chambray, denim, wool, silk, elastane, acrylic |

Table 3: Product properties from descriptions

### 4.4 Recommendation performance

#### 4.4.1 Comparison with baseline models

All the models except for MP were trained using stochastic gradient descent algorithm with random sampling. The results of MVBPR and baseline models were obtained on the full test set and the cold start test set with varied factor numbers. Latent factor was set to be 10 and total factor number were kept the same across all the baseline models. Experimental results were as shown in Table 4.

| #factor | MP | | MMMF | | BPRMF | | MVBPR | | improvement vs best | |
|---|---|---|---|---|---|---|---|---|---|---|
| | full | cold | full | cold | full | cold | full | cold | full | cold |
| 10 | 0.609 | 0.393 | 0.707 | 0.553 | 0.705 | 0.557 | **0.714** | **0.604** | 0.010 | 0.085 |
| 20 | 0.609 | 0.393 | 0.693 | 0.544 | 0.688 | 0.520 | **0.723** | **0.608** | 0.043 | 0.117 |
| 30 | 0.609 | 0.393 | 0.694 | 0.530 | 0.697 | 0.536 | **0.727** | **0.614** | 0.043 | 0.146 |
| 40 | 0.609 | 0.393 | 0.703 | 0.545 | 0.704 | 0.547 | **0.729** | **0.618** | 0.037 | 0.129 |
| 50 | 0.609 | 0.393 | 0.707 | 0.552 | 0.704 | 0.550 | **0.721** | **0.618** | 0.020 | 0.120 |
| 60 | 0.609 | 0.393 | 0.708 | 0.556 | 0.705 | 0.555 | **0.724** | **0.621** | 0.023 | 0.118 |
| 70 | 0.609 | 0.393 | 0.702 | 0.550 | 0.695 | 0.545 | **0.725** | **0.621** | 0.033 | 0.131 |
| 80 | 0.609 | 0.393 | 0.711 | 0.568 | 0.699 | 0.557 | **0.729** | **0.623** | 0.026 | 0.096 |
| 90 | 0.609 | 0.393 | 0.712 | 0.569 | 0.696 | 0.531 | **0.733** | **0.626** | 0.030 | 0.100 |
| 100 | 0.609 | 0.393 | 0.702 | 0.557 | 0.683 | 0.538 | **0.736** | **0.627** | 0.049 | 0.125 |
| 110 | 0.609 | 0.393 | 0.697 | 0.553 | 0.693 | 0.529 | **0.730** | **0.626** | 0.048 | 0.132 |
| 120 | 0.609 | 0.393 | 0.700 | 0.563 | 0.686 | 0.516 | **0.734** | **0.627** | 0.048 | 0.114 |
| 130 | 0.609 | 0.393 | 0.698 | 0.537 | 0.689 | 0.522 | **0.734** | **0.626** | 0.051 | 0.167 |
| 140 | 0.609 | 0.393 | 0.700 | 0.565 | 0.692 | 0.526 | **0.735** | **0.625** | 0.050 | 0.107 |
| 150 | 0.609 | 0.393 | 0.702 | 0.542 | 0.693 | 0.528 | **0.735** | **0.629** | 0.047 | 0.161 |

Table 4: AUC of MVBPR and baseline models

It can be easily observed that MP performed the worst as expected. BPRMF was slightly better than MMMF when factor number was small while MMMF gained some advantages as factor number increased. Through integration of visual and textual information, MVBPR outperformed all of the baseline models consistently. Moreover, MVPBR showed a superior performance in the cold-start case, achieving as high as 16% increase when factor number is 150. Furthermore, we could conclude from Table 5 that the performance advantages of MVBPR were significant compared with baseline models. The reason lies in that MVBPR considers various content information, which gives recommendation advice based on feedback as well as product related context, and this advantage is especially prominent in the cold start setting.

With factor number increasing from 10 to 150, BPRMF became stagnated in performance as AUC almost remained the same while MVBPR's performance continued to improve, i.e., implying a greater learning capacity of MVBPR. Similar phenomenon was observed on the cold start test set.

In addition, the dynamics of the training phase performance were also investigated. A plot of AUC trend was shown for different models as in Figure 4. The plot demonstrated that BPRMF and MMMF easily became overfitting and even got worse in the late period of training. Although MVBPR had a slower growth rate in the beginning, it showed great improvement potential as the training proceeded.

| Data | Hypothesis | t value | Sign. |
|---|---|---|---|
| Full | AUC (MVBPR) > AUC (MP) | 75.02 | *** |
| Full | AUC (MVBPR) > AUC (MMMF) | 11.20 | *** |
| Full | AUC (MVBPR) > AUC (BPRMF) | 10.57 | *** |
| Cold start | AUC (MVBPR) > AUC (MP) | 118.53 | *** |
| Cold start | AUC (MVBPR) > AUC (MMMF) | 23.18 | *** |
| Cold start | AUC (MVBPR) > AUC (BPRMF) | 17.91 | *** |

Table 5: Paired t-test of model performance

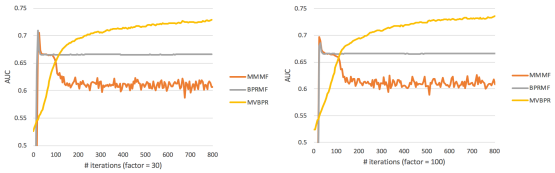


Figure 4: Performance dynamics in training phase

#### 4.4.2 Comparison on information sources

As a further step, this study also investigated how single models perform when only partial information is available. Experiments were conducted on different conditions and datasets. Figure 5 showed that both MVBPR-M (images only) and MVBPR-D (descriptions only) outperformed BPRMF while MVBPR-R (reviews only) performed worse than BPRMF, especially on the cold start test set. The reason could be that incorporating only reviews is not that informative for recommendation. In the cold start setting, since a product only had few reviews, this further constrained the expressive power of review features.

Compared with MVBPR-D that utilized descriptions, MVBPR-M was more effective in the cold-start setting, which showed the unique advantage of images. This could be explained by the fact that images are more persuasive and trustful in a cold start case, e.g., when consumers are confronted with a fresh new product.

To quantify the role of different information sources, models combining two sources of information, namely, M+D (image+description), M+R (image+review) and D+R (description+review) were built to compare their performance with the full model as shown in Figure 6.

Consistently, M+D performed better than M+R, and D+R had the least satisfactory performance.

In the last step, we quantified the performance improvement that each source could bring about. As demonstrated in Figure 7, images led to an improvement of 2.43% (full test set) and 10.42% (cold start test set), being the most prominent in the three sources. Reviews were the least important since it only contributed to a 0.36% and 0.08% increase in AUC performance. Descriptions stood between reviews and images, which had 2.01% and 2.15% improvement respectively. These results further confirmed that images indeed played an important role in recommendation, as the saying goes, "a picture is worth a thousand words".

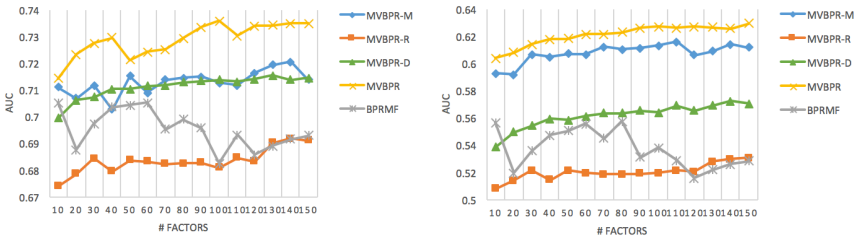


Figure 5: Performance comparison among single models (Left: full, Right: cold start)

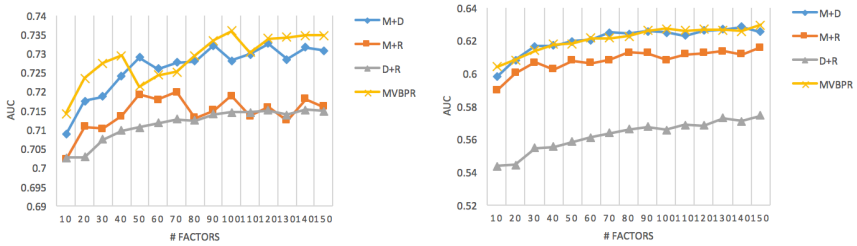


Figure 6: Model performance combining two information sources (Left: full, Right: cold start)

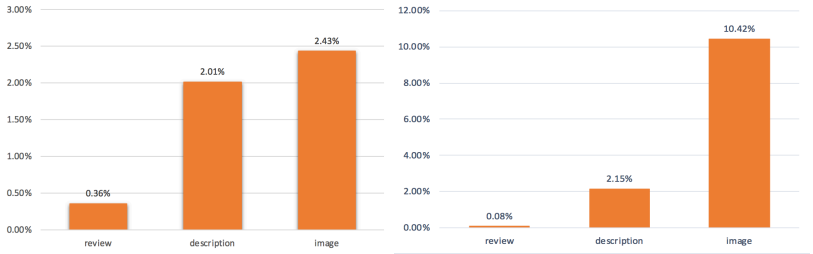


Figure 7: Performance improvement of each kind of information (Left: full, Right: cold start)

## 5 Conclusion and Future Work

This paper has proposed a Multi-View Bayesian Personalized Ranking (MVBPR) recommendation model that integrates implicit feedback, visual features and textual features into a matrix factorization framework. Effective techniques for information extraction including SCAE and LDA have been designed and the satisfactory performance was testified. In addition, extensive experiments showed the outperformance of the proposed model with an improvement as high as 16%. Moreover, this paper has also quantified the role of different information sources in the recommendation context. Images were far more useful than descriptions and reviews, further verifying the necessity of incorporating visual content in recommender system design.

Future work could be extended in two directions. One is to incorporate user generated images (in reviews or on social platforms) posted by consumers into the model, which are prevalent nowadays in marketing practices and may possess more fruitful and personalized information; the other direction is to explore a more general integration mechanism in other application contexts apart from online shopping considered in this study.

## Acknowledgments

The work was partly supported by the National Natural Science Foundation of China (71772101/71490724) and the MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (17JJD630006).

## References

[1] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state of the art and possible extensions, IEEE Transactions on Knowledge & Data Engineering 17 (2005) 734–749.

[2] D. Bawden, L. Robinson, The dark side of information: overload, anxiety and other

paradoxes and pathologies, Journal of Information Science 35(2) (2009) 180-191.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993-1022.
[4] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter? - An empirical investigation of panel data, Decision Support Systems 45 (2008) 1007-1016.
[5] Z. Gantner, S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, MyMediaLite: a free recommender system library, in: Proc. 5th ACM conference on Recommender Systems (RecSys'11), Chicago, US, 2011, pp. 305-308.
[6] A. Gogna, A. Majumdar, Matrix completion incorporating auxiliary information for recommender system design, Expert Systems with Applications 42(14) (2015) 5789-5799.
[7] Y Guan, Q Wei, G Chen, Deep learning based personalized recommendation with multi-view information integration, Decision Support Systems 118 (2019) 58-69.
[8] R. He, J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback, in: Proc. 13th AAAI Conference on Artificial Intelligence (AAAI'16), Phoenix, US, 2016, pp. 144-150.
[9] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, US, 2016, pp. 770-778.
[10] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42(8) 2009.
[11] A. Krizhevsky, Ii. Sulskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS'12), California, US, 2012, pp. 1097-1105.
[12] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521(7553) (2015) 436-444.
[13] E.J. Lee, S.Y. Shin, When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo, Computers in Human Behavior 31(2014) 356-366.
[14] K.H. Lim, I. Benbasat, L.M. Ward, The role of multimedia in changing first impression bias, Information Systems Research 11 (2) (2000) 115-136.
[15] H. Liu, J. He, T. Wang, W. Song, X. Du, Combining user preferences and user opinions for accurate recommendation, Electronic Commerce Research and Applications 12(1) (2013) 14–23.
[16] J. Liu, C. Wu, W. Liu, Bayesian probabilistic matrix factorization with social relations and item contents for recommendation, Decision Support Systems 55 (2013) 838–850.
[17] H. Ma, T.C. Zhou, M.R. Lyu, I. King, Improving recommender systems by incorporating social contextual information, ACM Transactions on Information Systems 29(2) (2011) 1–23.
[18] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction, in: Proc. International Conference on Artificial Neural Networks, Berlin, Germany, 2011, pp. 52–59.
[19] J. McAuley, J. Leskovec, Hidden factors and hidden topics: understanding rating dimensions with review text, in: Proc. 7th ACM Conference on Recommender Systems (RecSys'13), Hong Kong, China, 2013, pp. 165–172.
[20] L.A. Peracchio, J. Meyers-Levy, Using stylistic properties of ad pictures to communicate with consumers, Journal of Consumer Research 32(1) (2005) 29–40.
[21] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI'09), Montreal, Canada, 2009, pp. 452-461.
[22] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. 6th International Conference on Learning Representations (ICLR'15), San Diego, US, 2015.
[23] C. Wang, D.M. Blei, Collaborative topic modeling for recommending scientific articles, in: Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11), San Diego, US, 2011, pp. 448-456.
[24] M. Wang, X. Li, P.Y. Chau, The impact of photo aesthetics on online consumer shopping behavior: an image-processing-enabled empirical study, in: Proc. International Conference on Information Systems (ICIS'16), Dublin, Ireland, 2016.
[25] S. Wang, Y. Wang, J. Tang, K. Shu, S, What your images reveal: Exploiting visual contents for point-of-interest recommendation, in： Proc. 26th International Conference on World Wide Web (WWW'17), Perth, Australia, 2017, pp. 391-400.
[26] H.Wang, N. Wang, D.Y. Yeung, Collaborative deep learning for recommender systems, in Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15), Sydney, Australia, 2015, pp. 1235-1244.
[27] F. Zhang, N.J. Yuan, D. Lian, X. Xie, W. Ma, Collaborative knowledge base embedding for recommender systems, in: Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), San Francisco, US, 2016, pp. 353-362.
[28] F. Zhu, X. Zhang, Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics, Journal of Marketing 74(2) (2010) 133-148.