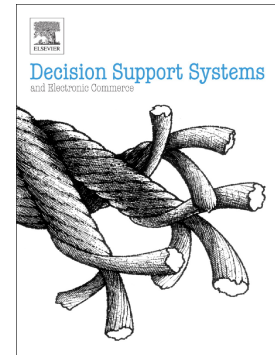


Accepted Manuscript

Deep learning based personalized recommendation with multi-view information integration

Yue Guan, Qiang Wei, Guoqing Chen



PII: S0167-9236(19)30011-9
DOI: <https://doi.org/10.1016/j.dss.2019.01.003>
Reference: DECSUP 13024
To appear in: *Decision Support Systems*
Received date: 1 August 2018
Revised date: 17 January 2019
Accepted date: 17 January 2019

Please cite this article as: Yue Guan, Qiang Wei, Guoqing Chen , Deep learning based personalized recommendation with multi-view information integration. Decsup (2019), <https://doi.org/10.1016/j.dss.2019.01.003>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Deep Learning Based Personalized Recommendation with Multi-View Information Integration

Yue Guan, Qiang Wei¹, Guoqing Chen

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

Abstract

With the rapid proliferation of images on e-commerce platforms today, embracing and integrating versatile information sources have become increasingly important in recommender systems. Owing to the heterogeneity in information sources and consumers, it is necessary and meaningful to consider the potential synergy between visual and textual content as well as consumers' different cognitive styles. This paper proposes a multi-view model, namely, Deep Multi-view Information iNtEgration (Deep-MINE), to take multiple sources of content (i.e., product images, descriptions and review texts) into account and design an end-to-end recommendation model. In doing so, stacked auto-encoder networks are deployed to map multi-view information into a unified latent space, a cognition layer is added to depict consumers' heterogeneous cognition styles and an integration module is introduced to reflect the interaction of multi-view latent representations. Extensive experiments on real world data demonstrate that Deep-MINE achieves high accuracy in product ranking, especially in the cold-start case. In addition, Deep-MINE is able to boost overall model performance compared with models taking a single view, further verifying the proposed model's effectiveness on information integration.

Keywords:

Multi-view information; Deep learning; Information integration; Personalized recommendation; Representation learning

¹ Corresponding author, weiq@sem.tsinghua.edu.cn

1. Introduction

The amount and variety of data are increasing exponentially in today's online marketplaces, of which multimedia and user-generated content account for a large proportion². While consumers benefit from such rich and helpful information, they also face an information overload problem in their online shopping processes [7][22]. To alleviate this problem, some technological tools have been developed to assist consumers in product search and decision-making. Recommender systems are one of the most widely applied decision aids that aim to provide personalized recommendation services for consumers. Recommender systems conduct in-depth mining of historical records, infer consumer preferences from the data and recommend products that consumers may like [1].

Plenty of efforts have been devoted to investigating different types of data, e.g., consumers' profile, online reviews, product descriptions, and social network to enhance recommendations [12],[33],[36],[48], whereas images have not been comprehensively utilized yet because of the complexity of image processing and the difficulty of integrating images into recommendations. However, over the past few years, the advancement of deep learning techniques has made it possible to have a deeper understanding of multimedia content, in which some studies attempt to extract valuable information from visual data through deep neural networks [14][27][51].

For online consumers, their first impressions are usually derived from the visual appeal of products [23]. The perceptual and persuasive advantages of images have been well demonstrated in consumer behavior research [43]. Product images provide us with visual cues, reduce perceived risk, possess high attention-grabbing qualities and are remembered better [9][54]. Furthermore, when consumers make purchase decisions online, they naturally consider multiple sources of information together, e.g., images, descriptions, and reviews, meaning that they take a multi-view perspective.

Concerning visual and textual content, prior studies have found that visual messages are complements, but not alternatives to textual content [31]. For instance, a dress may have a description like "floral printed, round neck, long sleeve, two-side pockets". Though helpful, the description does not provide details of floral pattern and where exactly the pockets are located, which a consumer prefers to know. Meanwhile, an image could clearly show the relevant information to fill the

² <https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/>

information gap, which is particularly desirable for online shoppers targeting experience goods, such as clothes. However, some features (e.g., fabrics) can only be accurately described in text and are unlikely to be inferred from images. Therefore, it is deemed intuitive and meaningful to combine image and textual content from such a multi-view perspective to effectively enhance the quality of recommendations.

Apart from the information heterogeneity caused by images and texts, there is also heterogeneity in consumers due to their differences in cognitive styles. Cognitive style refers to the way people think, perceive and memorize information, which significantly influences people's behavior and decision-making process [29][49]. There are some studies discussing the cognitive style model, of which the Verbal-Imagery dimension in Cognitive Style Analysis in [45] is most closely related to our research. The Verbal-Imagery dimension describes individual's mode of information representation in memory during thinking. Concretely, affective users are more sensitive to visual content, while cognitive users are more sensitive to verbal clues [50]. Multi-view-based recommendation is expected to improve further by incorporating cognitive style heterogeneity and information heterogeneity.

Leveraging the heterogeneities in product content and users (users and consumers are used interchangeably in the following content) in recommender system design involves a twofold challenge. First, the heterogeneity in product content requires different modeling techniques to ensure an appropriate representation for each view of content, and the heterogeneity in users requires reflecting different attentive preferences of users on the respective content. Second, an overall mechanism needs to be developed to formulate a unified representation that embraces various representations for the respective content with users' diversified cognition styles. Note that, although some attempts have discussed the complementary relationship between visual and textual views in a general manner [31], the challenge has not been adequately addressed, which motivates our work.

This paper proposes a multi-view recommendation model, namely, Deep Multi-view Information Integration (Deep-MINE), where visual and textual content are leveraged with representation learning techniques and mapped into a unified latent space. Furthermore, a cognition factor is introduced to characterize the heterogeneity in users' cognitive styles. Then, the embedding approach is deployed to automatically learn the interaction between visual and textual content enhanced with

individualized cognitive styles. The effectiveness of Deep-MINE is verified through extensive experiments. It is also worth mentioning that, Deep-MINE shows merit in dealing with the cold-start problem to some extent by taking a comprehensive multi-view perspective.

The rest of the paper is organized as follows. Section 2 reviews the related literature. Section 3 presents the model framework and formulation, as well as the parameter learning process and recommendation procedure. The data experiments in Section 4 demonstrates the outperformance of the proposed model over baselines. The conclusion and future work are presented in Section 5.

2. Related work

2.1 Recommender systems

Generally, recommendation models can be grouped into three categories: content-based models, collaborative filtering (CF) and hybrid models [1]. Content-based models recommend items similar to those users have liked previously based on item or user characteristics. CF models recommend items according to the similarities among users or items. Matrix factorization (MF) [24] is an effective CF-based method. MF decomposes the feedback matrix into two low-dimension matrices, i.e., the item latent factor matrix and user latent factor matrix, and the interaction between the two matrices represents the preference score. However, it suffers from the cold start problem, as the latent factors can hardly be inferred, if there is no historical feedback available. Hybrid models that combine the two methods above have been widely used, which consider content information as well as collaborative preferences. The content information includes the user profile, item descriptions, social relations and social network [12],[33],[36],[48]. As an important source of user-generated content, review texts are also utilized to elicit user preferences. For instance, HFT proposes to combine latent rating dimensions with latent review topics learned by topic models to make latent factors more interpretable [38]. CTR recommends scientific articles with similar ideas in online researcher communities [52]. Liu et al. [32] extract consumer opinions from reviews with aspect-based opinion mining and make recommendations based on extracted opinions.

Targeting recommendation objectives, recommender systems could be divided into point-wise recommendation and pairwise recommendation. Point-wise recommendation was widely used in the

early days such as for movie recommendations, aiming to predict the rating or score for each user-item pair. By contrast, pairwise recommendation aims to optimize the ranking for potential candidates rather than focusing on the absolute rating scores, which is more realistic. One of the most influential studies is the work by Rendle et al. [44], which proposes a generalized Bayesian Personalized Ranking (BPR) framework and has been widely applied in top-n recommendation [58], session based recommendation [17], group-based recommendation [42] and point-of-interest recommendation [8].

However, existing efforts did not satisfactorily incorporate visual information, such as images. Furthermore, owing to the complexity of image processing, how to organically integrate images with other information to facilitate recommendation needs to be explored.

2.2 Image-aware recommendation

As mentioned earlier, images are influential and necessary in consumer decision-making on e-commerce platforms. There are some studies that apply visual signals to item recommendations in the field of computer science. In the early days, image-based recommendations mainly focus on image retrieval with feature engineering [5][21][35]. Considering the wide acceptance of deep learning in industrial and academic fields, some studies manage to leverage deep learning techniques to take advantage of image and textual information in recommendation models [57]. Wang et al [53] prove that deep learning-based models outperform traditional topic-based models. Some representative models include VBPR [15], VPOI [55] and CKE [56]. VBPR represents each image with a 4096-dimension feature vector, which is extracted from a pre-trained image classification model [26] and adds an embedding layer on top of it to obtain a dense item representation. Similarly, Wang et al. [55] proposes a POI recommendation model that utilizes a pre-trained VGG-16 model. Nevertheless, due to the differences of image contexts, a pre-trained model in a general-purpose image classification task may not well fit specific recommendation tasks. Moreover, there is usually a lack of well-recognized and predefined labels for images, therefore supervised models are not applicable either.

There are also some customized deep neural networks for image-aware recommendations. Lei et al.

[28] propose a dual-net deep network to map the images and preferences of users into the same latent semantic space. Deepstyle [34] considers style features as well as category information of item to fully account for visual signals. CKE [56] proposes a collaborative knowledge embedding model that leverages image, text and structural information in a single Bayesian model. Visual embedding and textual embedding are implemented through two auto-encoder structures, i.e., stacked convolutional auto-encoder (SCAE) and stacked denoising auto-encoder (SDAE). However, the relationship between different types of knowledge is not well addressed, and the textual content originate from external knowledge bases, which is generally inaccessible for the online shopping recommendation context. In addition, some conceptual multi-view frameworks are highlighted in consumer behavioral and psychological research [20][30][31], which however lack technical treatments.

2.3 Representation learning

Representation learning is an effective tool that has been widely used in machine learning. The key idea of representation learning is to seek a low-dimensional embedding of data while preserving different discriminative factors of variation behind data. Several kinds of neural networks have been proposed to extract features from unstructured data, including undirected models such as the Deep Bayesian network (DBN) [18], Restricted Boltzmann Machine (RBM) [47], and directed models, such as the auto-encoder [4]. A stacked auto-encoder [41] is a kind of unsupervised model with multiple layers of auto-encoders in which the output of each layer is wired to the input of the successive layer. Aiming to reconstruct the original input as well as to compress the original high dimension input, it is composed of encoder and decoder parts. To constrain the representation from duplicating the input, auto-encoders are usually regularized and several variants of auto-encoders are proposed, including Contractive Auto-encoders (CAE) and Denoising Auto-encoder (DAE) [53][56]. For images, the Convolutional Neural Network (CNN) shows considerable advantages as it preserves the input's neighborhood relations and spatial locality in their latent higher-level feature representations [27]. Thus the convolutional stacked auto-encoder is a natural choice for image feature representation in this study.

The embedding approach is also widely adopted for information representation. To overcome the

adaptation problem, an embedding layer is usually imposed on top of features extracted from pre-trained deep learning models to obtain a dense feature representation [15][55]. Based on previous achievements, this study adopts the auto-encoder structure to obtain a latent representation for each view of information and designs an embedding approach to exploring the interaction of multi-view features.

2.4 Cognitive styles

In the research field of psychology and education, a variety of research studies have discussed the constructs, theories and models related to cognitive styles [25]. Messick [40] defines cognitive style as stable attitudes, preferences or habitual strategies that determine individuals' modes of perceiving, remembering, thinking, and problem solving. Cognitive style has been widely applied in personnel selection, career guidance, task design, team composition, and conflict management [2][6]. Nevertheless, to our knowledge, few studies consider users' different cognitive styles on an e-commerce platform.

According to Riding and Cheema [45], an individual's cognitive style can be positioned on two orthogonal dimensions, namely, Wholist-Analytic and Verbal-Imagery. The Verbal-Imagery dimension describes individuals' mode of information representation in memory. Verbalizers are those who tend to process information in words, and they learn better from textual input, while visualizers learn better from pictorial presentation [46]. A user's position on this dimension is of critical importance in deciding the relative weights of image and textual content in online purchase decisions.

Previous measurements of cognitive styles mainly focus on self-report measures, which may not be effective in certain cases due to the questionable reliability and validity [10]. Furthermore, on a real online-shopping platform, hundreds of thousands of consumers browse product information, make purchase decisions and write product reviews from time to time. Hence, from an operational point of view, it is extremely difficult to explicitly assess these consumers' cognitive styles through the traditional measurement of cognitive style. In response to the call for utilizing multiple methods [3] and based on the similar idea in [13] where online user's cognitive styles are inferred in a Bayesian learning process through each user's clickstream data, this paper proposes a data-driven measurement

to learn an individual's cognitive style through observations of a consumer's historical purchase behavior. Furthermore, the extracted individualized cognitive styles are incorporated to facilitate personalized recommendations, which contributes to the field of recommender systems.

3. Model framework and computational methods

3.1 Problem formulation

Focusing on recommendation in the online shopping setting, in a multi-view information context, let J be the set of all items concerned; for each item j , it has at least one image M_j , a description D_j , and a set of reviews $R_{j1}, R_{j2}, \dots, R_{jm}$. Let I be the set of all users; for each specific user i , his/her purchase history is known, and all users' purchase histories constitute an adjacency matrix X , in which $X_{ij} = 1$ means that user i purchased item j , and $X_{ij} = 0$ otherwise.

Generally, a user purchase can be treated as a kind of implicit feedback [19], as it indirectly reflects the user's preference. Without loss of generality, assume user i bought item j instead of item j' ($j, j' \in J$), then user i implicitly prefers j to j' , denoted as $j >_i j'$ [15]. Consistently with [44], item pairs are used as training data in this paper. More specifically, suppose that for user i , the set of all the items he/she bought is denoted as J_i^+ ($J_i^+ \subset J$), the data set can be formalized as $S = \{(i, j, j') | j \in J_i^+, j' \in J - J_i^+\}$, or equally, $S = \{(i, j, j') | X_{ij} = 1, X_{ij'} = 0\}$. Therefore, the recommendation task is to derive a personalized list for each user based on those items which he/she has not provided any feedback.

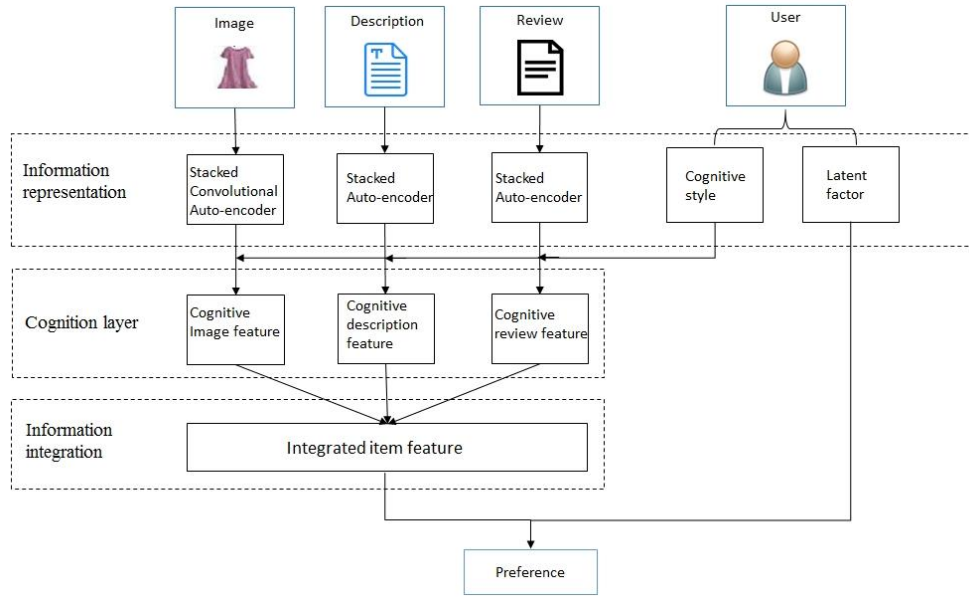


Fig. 1. Deep-MINE Model Framework

3.2 Deep-MINE model

The overall model of Deep MINE consists of three parts: information representation, cognition layer and information integration. The model framework is as shown in Fig. 1.

3.2.1 Multi-view information representation

This subsection aims to map heterogeneous information into a unified latent space, in which a latent factor representing the source information is obtained through a deep neural network.

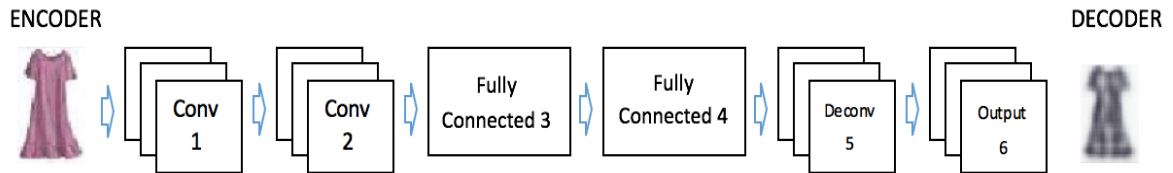


Fig. 2. Stacked Convolutional Auto-encoder for Images

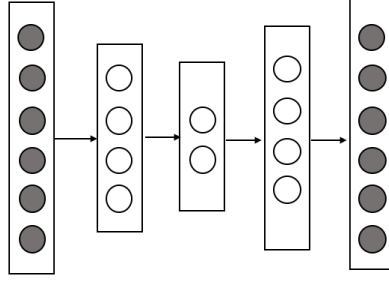


Fig. 3. Stacked Auto-encoder for Texts

For images, a 6-layer stacked convolutional auto-encoder network [37] is designed. On the one hand, convolutional networks could preserve the input's neighborhood relations and spatial locality in their latent higher-level feature representations and show a superior performance in image classification related tasks [14][51]. On the other hand, the auto-encoder structure could preserve as much discriminative information as possible. For unstructured texts, a bag-of-word representation for each item is obtained and a 4-layer stacked auto-encoder network is designed to obtain its latent representation through layer-wise dimension reduction [41][53]. The number of layers for the auto-encoder networks is consistent with previous literature [53][56], and some empirical results for different number of layers are discussed in Section 4.3.5. Fig. 2 and Fig. 3 show the two types of auto-encoder networks, and the generation process of latent representations is detailed as follows.

For stacked convolutional auto-encoder, layers 1, 2, 5, and 6 are convolutional layers and layers 3 and 4 are fully connected layers. Suppose that an input image of item j is denoted as M_0 . For each layer l , let each column k of its weight parameters W_l follows a normal distribution, namely, $W_{lk} \sim N(0, \lambda_w^{-1}I)$, and let bias parameters $b_l \sim N(0, \lambda_b^{-1}I)$, where I refers to the identity matrix. If $l = 1, 2, 5, 6$, then the output of each layer l depends on the convolution operation of the network parameters and output of the last layer, namely, $M_l = \sigma(W_l * M_{l-1} + b_l)$, where $*$ represents convolution operation. If $l = 3, 4$, then $M_l = \sigma(W_l \cdot M_{l-1} + b_l)$, where \cdot represents matrix multiplication. The middle layer output M_3 is used as the visual feature representation for item j . As implied by the name *auto-encoder*, the input is reconstructed at the last layer. To summarize, the encoder network and decoder, sharing the same weight matrices, are represented as Eq. (1) Eq. (2),

respectively.

$$M_3 = g_1(M_0, W, b) = \sigma(W_3 \cdot \sigma(W_2 * \sigma(W_1 * M_0 + b_1) + b_2) + b_3) \quad (1)$$

$$M_6 = g'_1(M_3, W, b) = \sigma(W'_1 * \sigma(W'_2 * \sigma(W'_3 \cdot M_3 + b_4) + b_5) + b_6) \quad (2)$$

For the stacked auto-encoder, suppose the textual description of item j is denoted as D_0 . For each layer l , let each column k of its weight parameters Q_l follow a normal distribution, namely $Q_{lk} \sim N(0, \lambda_q^{-1} I)$, and let bias parameters $c_l \sim N(0, \lambda_c^{-1} I)$, where I is the identity matrix. The output of each layer l depends on the weight matrix, bias parameters and the output of the last layer, namely, $D_l = \sigma(Q_l \cdot D_{l-1} + c_l)$. The middle layer output D_2 is used as the textual representation for item j . Therefore, the encoder network and the decoder are represented as Eq. (3) and Eq. (4), respectively.

$$D_2 = g_2(D_0, Q, c) = \sigma(Q_2 \cdot \sigma(Q_1 \cdot D_0 + c_1) + c_2) \quad (3)$$

$$D_4 = g'_2(D_2, Q, c) = \sigma(Q'_1 \cdot \sigma(Q'_2 \cdot D_2 + c_3) + c_4) \quad (4)$$

Note that, as reviews are also textual, a similar 4-layer stacked auto-encoder is built for R_0 as that for D_0 , with $R_2 = g_3(R_0, N, t)$, $R_4 = g'_3(R_2, N, t)$. Furthermore, to ensure that we obtain an effective representation for each view of information, the Mean Square Error (MSE) loss functions L_1, L_2, L_3 are introduced with the aim of minimizing the reconstruction error, and regularization terms are also added to control the magnitude of the network parameters (i.e., Eq. (5) - (7)), where $\lambda_m, \lambda_d, \lambda_r$ are hyperparameters and $\lambda_w, \lambda_b, \lambda_q, \lambda_c, \lambda_n, \lambda_t$ are parameters of the corresponding normal distributions.

$$\min L_1 = \frac{\lambda_m}{2} \sum_j \|M_L - M_0\|_2^2 + \frac{\lambda_w}{2} \sum_l \|W_l\|_2^2 + \frac{\lambda_b}{2} \sum_l \|b_l\|_2^2 \quad (5)$$

$$\min L_2 = \frac{\lambda_d}{2} \sum_j \|D_L - D_0\|_2^2 + \frac{\lambda_q}{2} \sum_l \|Q_l\|_2^2 + \frac{\lambda_c}{2} \sum_l \|c_l\|_2^2 \quad (6)$$

$$\min L_3 = \frac{\lambda_r}{2} \sum_j \|R_L - R_0\|_2^2 + \frac{\lambda_n}{2} \sum_l \|N_l\|_2^2 + \frac{\lambda_t}{2} \sum_l \|t_l\|_2^2 \quad (7)$$

3.2.2 Cognition layer

As mentioned above, according to the Verbal-Imagery dimension of cognitive style, users are heterogeneous in information processing, implying that some users may value images more, while others may pay more attention to texts. In [10], cognitive style is obtained from lab experiments where participants are asked to complete a survey on certain tasks. In a real online shopping environment, however, the consumers are usually in hundreds of thousands, i.e., it is extremely difficult to explicitly assess their cognitive styles through the above methods. Therefore, we propose an integrated model to learn their cognitive styles in a data-driven fashion from users' implicit feedback, i.e., purchase information. In this spirit, a cognition layer is added in the Deep-MINE model architecture. Specifically, a 3-dimension vector is imposed on three different views, i.e., images, descriptions and reviews, to represent an individual's cognitive style, which is a reasonable extension on Verbal-Imagery dimension, where a textual description is separated from a textual review, because they are quite different in terms of the content, position and form of presentation in the e-commerce context.

Concretely, user i 's cognition factor is denoted as $[a_{i1}, a_{i2}, a_{i3}]$. From Section 3.2.1, as the latent representations for the three types of information are denoted as M_3, D_2, R_2 , the perceived information for user i is moderated as $[a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2]$. The value of the cognition factor is to be learned during the model training phase.

3.2.3 Multi-view information integration

With user perceived information available, an integration module is proposed to build a full picture of an item from a multi-view perspective. First, a stacking layer (concatenation) of different representations is formulated as Eq. (8), where $c(\cdot)$ represents the concatenation. Then, an embedding layer is deployed to transform the concatenated factor into a lower dimension factor as Eq. (9).

$$f^c = c(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2) \quad (8)$$

$$f_j = W_{fu} \cdot f^c \quad (9)$$

Notably, this embedding operation is deemed to be a key step to enable reorganizing and utilizing available information in the following steps [15]. Similar to the cognition factor, the specific weights in W_{fu} are not known to us beforehand and need to be learned in the model training phase, which is natural, as the integration mechanism is highly dependent on the specific context. So far, f_j can be deemed as content factor, as all three pieces of information have been integrated together.

Apart from images, descriptions and reviews, there may exist additional information about the item outside the e-commerce platform, which could potentially affect the consumer purchase behavior. Therefore, factor v_j is introduced to capture the hidden information. Eventually, the aggregated item factor $item_j$ consisting of hidden information and content factor is deemed to represent the item comprehensively, which is shown in Eq. (10).

$$item_j = c(v_j, f_j) \quad (10)$$

3.2.4 User preference

Based on previous discussions, for recommendation purposes, the preference x_{ij} of user i on item j can be formulated as Eq. (11), where v_j and f_j represent the hidden information and content factor of an item as introduced above, u_i and θ_i are the user perception factors corresponding to v_j and f_j , and α_i and β_j denote user bias and item bias, respectively.

$$x_{ij} = \alpha_i + \beta_j + u_i^T v_j + \theta_i^T f_j \quad (11)$$

To combine all the parts above, the preference of user i on item j could be formulated as Eq. (12). Consistently with [44], the probability of user i preferring item j to item j' can be formulated with a sigmoid function as Eq. (13).

$$x_{ij} = \alpha_i + \beta_j + u_i^T v_j + \theta_i^T (W_{fu} \cdot c(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2)) \quad (12)$$

$$P(j >_i j') = \sigma(x_{ij} - x_{ij'}) = 1 / (1 + \exp(-(x_{ij} - x_{ij'}))) \quad (13)$$

3.2.5 Objective function

To optimize the whole model and learn the model parameters, an objective function is formulated in this subsection. As Deep-MINE has two major parts, namely, representation learning and preference learning, the objective function performs two tasks. One task is to maximize the logarithm of the ranking probability, i.e., $\sum_{(i,j,j') \in S} \ln \sigma(x_{ij} - x_{ij'})$. The other task is to obtain an effective information representation. Hence the loss functions of the auto-encoder networks, i.e., $L_1 + L_2 + L_3$, in information representation layer need to be included. In addition, regularization terms for related model parameters are added to avoid overfitting. Overall, the objective function can be formulated as Eq. (14), where S is the training set consisting of triple (i, j, j') as explained in Section 3.1. $\lambda_m, \lambda_d, \lambda_r, \lambda_\theta, \lambda_\beta, \lambda_{W_{fu}}$ are hyperparameters controlling the relative weights of different components in the objective function.

$$\begin{aligned} \max \mathcal{L}(W, b, Q, c, N, t, \beta, u, v, \theta, \alpha, W_{fu}) = & \sum_{(i,j,j') \in S} \ln \sigma(x_{ij} - x_{ij'}) - \frac{\lambda_m}{2} \sum_j \|M_L - M_0\|_2^2 - \\ & \frac{\lambda_d}{2} \sum_j \|D_L - D_0\|_2^2 - \frac{\lambda_r}{2} \sum_j \|R_L - R_0\|_2^2 - \frac{\lambda_\theta}{2} \sum_i \|u_i\|_2^2 - \frac{\lambda_\theta}{2} \sum_j \|v_j\|_2^2 - \frac{\lambda_\theta}{2} \sum_i \|\theta_i\|_2^2 - \frac{\lambda_\beta}{2} \sum_i \|\beta_i\|_2^2 - \\ & \frac{\lambda_{W_{fu}}}{2} \|W_{fu}\|_2^2 - (\frac{\lambda_w}{2} \sum_l \|W_l\|_2^2 + \frac{\lambda_b}{2} \sum_l \|b_l\|_2^2) - (\frac{\lambda_q}{2} \sum_l \|Q_l\|_2^2 + \frac{\lambda_c}{2} \sum_l \|c_l\|_2^2) - (\frac{\lambda_n}{2} \sum_l \|N_l\|_2^2 + \\ & \frac{\lambda_t}{2} \sum_l \|t_l\|_2^2) \end{aligned} \quad (14)$$

Considering the complexity and nonlinear relationships of parameters, it is impossible to find a closed form solution [56]. An iterative algorithm is developed, as discussed in the next subsection.

3.3 Parameter learning

As the objective function is based on pairwise items, for each user, there are many more items for

which he/she has not provided any feedback (i.e., purchase) than those that he/she has purchased. Therefore, a negative sampling strategy [56] is adopted to randomly sample one item pair from the training set S each time and update the corresponding parameters with stochastic gradient descent.

To obtain the gradient with respect to each parameter, back propagation is used during each update. The update formulas for the parameters are shown in Eqs. (15)-(22), where $x_{ijj'} = -(x_{ij} - x_{ij'})$, lr denotes the learning rate, t denotes the batch number. $\frac{\partial f(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2)}{\partial (a_{i1} \cdot M_3)}$ is a sparse matrix with the first n rows being an identity matrix and the remaining $m + k$ rows being zeros, where n, m, k are the latent factor dimensions of images, descriptions and reviews. Both $\frac{\partial g_1(M_0, W, b)}{\partial W_l}$ and $\frac{\partial (g_1'(g_1(M_0, W, b)))}{\partial W_l}$ can be directly derived with back propagation, which are not expanded here for simplicity. The update formulas for N, Q, c, t are omitted, as they could be derived in a similar fashion as W, b . The algorithmic details of parameter learning process are provided in Algorithm 1.

$$\beta_i^{t+1} = \beta_i^t + lr \cdot (\sigma(x_{ijj'})) \quad (15)$$

$$\theta_i^{t+1} = \theta_i^t + lr \cdot (\sigma(x_{ijj'})) W_{fu}^t (f_j^c - f_{j'}^c) - \lambda_\theta \theta_i^t \quad (16)$$

$$v_j^{t+1} = v_j^t + lr \cdot (\sigma(x_{ijj'})) \cdot u_i^t - \lambda_\theta v_j^t \quad (17)$$

$$u_i^{t+1} = u_i^t + lr \cdot (\sigma(x_{ijj'})) \cdot (v_j^t - v_{j'}^t) - \lambda_\theta u_i^t \quad (18)$$

$$W_{fu}^{t+1} = W_{fu}^t + lr \cdot (\sigma(x_{ijj'})) \cdot \theta_i^t \cdot (f_j^{cT} - f_{j'}^{cT}) - \lambda_{W_{fu}} W_{fu}^t \quad (19)$$

$$a_{i1}^{t+1} = a_{i1}^t + lr \cdot \sigma(x_{ijj'}) \cdot \theta_i^t \cdot W_{fu}^t \cdot \frac{\partial f(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2)}{\partial (a_{i1} \cdot M_3)} \cdot g_1(M_0, W, b) \quad (20)$$

$$\begin{aligned} W_l^{t+1} = & W_l^t + lr \cdot (\sigma(x_{ijj'})) \cdot (\theta_i^t)^T \cdot W_{fu}^t \cdot \frac{\partial f(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2)}{\partial (a_{i1} \cdot M_3)} \cdot a_{i1}^t \cdot \left(\frac{\partial g_1(M_0, W, b)}{\partial W_l} - \frac{\partial g_1(M_{f0}, W, b)}{\partial W_l} \right) - \\ & \lambda_m (M_6 - M_0) \frac{\partial (g_1'(g_1(M_0, W, b)))}{\partial W_l} - \lambda_w \cdot W_l^t \end{aligned} \quad (21)$$

$$\begin{aligned} b_l^{t+1} = & b_l^t + lr \cdot (\sigma(x_{ijj'})) \cdot (\theta_i^t)^T \cdot W_{fu}^t \cdot \frac{\partial f(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2)}{\partial (a_{i1} \cdot M_3)} \cdot a_{i1}^t \cdot \left(\frac{\partial g_1(M_0, W, b)}{\partial b_l} - \frac{\partial g_1(M_{f0}, W, b)}{\partial b_l} \right) - \\ & \lambda_m (M_6 - M_0) \frac{\partial (g_1'(g_1(M_0, W, b)))}{\partial b_l} - \lambda_b \cdot b_l^t \end{aligned} \quad (22)$$

Algorithm 1. Deep-MINE parameter learning.

Input: S, I, J, M, D, R

Output: $W, b, Q, c, N, t, \beta, u, v, \theta, a, W_{fu}$

Initialize the parameters $W, b, Q, c, N, t, \beta, u, v, \theta, a, W_{fu}$

WHILE $epoch \leq training_epoch$

WHILE $batch \leq total_batch$

- a) Randomly choose an item pair (i, j, j') from S .
- b) Extract corresponding images, descriptions and review texts of j and j' as M_0, D_0, R_0 and $M_{j'0}, D_{j'0}, R_{j'0}$
- c) Forward phase:
 - i. Forward the input through the information representation module and generate visual and textual features $M_3 = g_1(M_0, W, b), D_2 = g_2(D_0, Q, c), R_2 = g_3(R_0, N, t)$.
 - ii. Forward the features through the user cognition layer and obtain perceived information $a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2$
 - iii. Forward through the information integration module and obtain the item factor $item_j = c(v_j, W_{fu} \cdot c(a_{i1} \cdot M_3, a_{i2} \cdot D_2, a_{i3} \cdot R_2))$
 - iv. Calculate the user preference score and probability
$$x_{ij} = \alpha_i + \beta_j + u_i v_j + \theta_i f_j$$

$$x_{ij'} = \alpha_i + \beta_{j'} + u_i v_{j'} + \theta_i f_{j'}$$

$$P(j >_i j') = \sigma(x_{ij} - x_{ij'})$$
- d) Backward phase:
 - i. Obtain the objective function $\mathcal{L}(W, b, Q, c, N, t, \beta, u, v, \theta, a, W_{fu})$ as Eq. (14)
 - ii. Calculate gradients and update parameters using Eqs. (15)-(22).

END WHILE

IF $|\mathcal{L}_{epoch} - \mathcal{L}_{epoch-1}| < \delta$, *THEN*

BREAK

END IF

END WHILE

3.4 Prediction and recommendation

An online recommendation for user i can be conducted as follows. For each item j , following the Deep-MINE framework, given its content information and user characteristics, the preference x_{ij} could be assessed using Eq. (12). Then, a sorting operation is performed for x_{ij} s, and the top K items that constitute the recommendation list are selected. The framework is illustrated in Fig. 4.

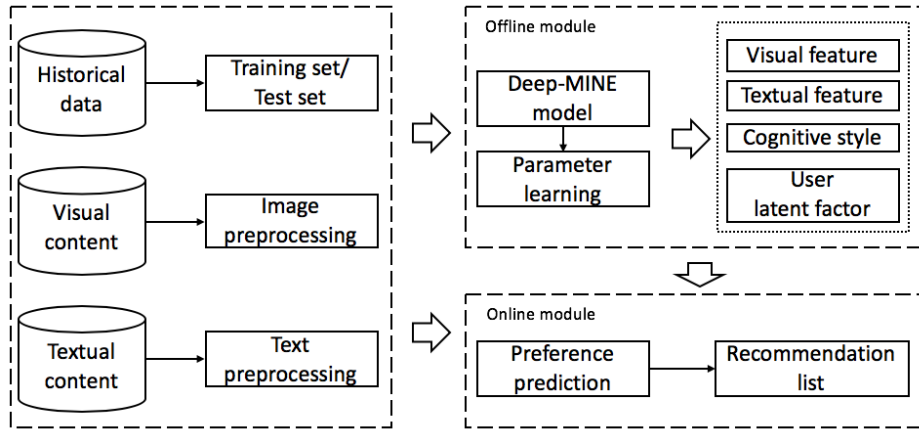


Fig. 4. Framework for Deep-MINE recommender system

4 Empirical study and results


4.1 Data description

To demonstrate the effectiveness of the proposed model, two categories of real-world datasets were obtained from Amazon.com, namely, Women's Dresses and Baby Clothes. Both Women's Dresses and Baby Clothes are typical experience goods, for which images and reviews are well recognized to be highly informative, complementing the product descriptions for consumer decision-making. For the Baby Clothes dataset, only image features are available (<http://jmcauley.ucsd.edu/data/amazon/>) [16][39]. Thus, the image auto-encoder network is applied to the Women's Dresses dataset.

The data preprocessing is conducted as follows. As a product usually had 3–4 images to show it from different angles, without loss of generality, one image was randomly chosen as the input. For textual content, a bag-of-words approach was used and words with high frequency were kept in the vocabulary. To control the dimension of the input data and prevent the negative impact of misspelled words or typos, words that appeared less than 10 times were deleted for the Women's Dresses dataset and words that appeared less than 100 times were deleted for the Baby Clothes dataset (the corpus of the Baby Clothes dataset is much larger than that of the Women's Clothes dataset, which explains the different thresholds), which is consistent with [52][53]. A stop words list was also kept to delete words, such as *of* and *in*. The text preprocessing steps included capital words conversion, word stemming and stop words deletion. Finally, 1,461 and 1,894 words were retained for the descriptions

and reviews in the Women's Dresses dataset. 1,516 and 1,502 words were retained for the Baby Clothes dataset. Note that, most consumers would not read all of the reviews and reviews with helpful votes were usually displayed with priority by the platform. Hence, we only kept the reviews with helpful votes for preprocessing. An example of product content is presented in Table 1.

Table 1. An Example of Product Content

Image	Description	Review
	<i>95% Polyester 5% Spandex. Hand wash; dry flat; Model Wearing Size 1X. Height: 5'9" Waist:37.5" Hips: 42" Bust:34" Super soft fabric defines this curve-loving dress, while a surplice neckline and billowing silhouette add gentle whimsy.</i>	<i>Looks just like Brigitte Bailey and Lovestitch style but the material was really thin and did not drape as nicely. The multi blue color is attractive and the price was good but I'm spoiled on the other two brands. And will stick with those.</i>

For both datasets, we ensured that each consumer had at least two feedbacks, i.e., one for the training set and one for the test set, as the proposed model needs to infer consumer preferences from one's purchase history. All products were kept in the dataset, including products that had few or no feedbacks (i.e., new products released to the market), which are known as cold start products. Initially, the Women's Dresses dataset had 256,749 feedbacks and the Baby Clothes dataset had 32,419 feedbacks. After data cleaning and preprocessing, we finally had 5,981 users and 2,579 products for the Women's Dresses dataset and 8,018 users and 3,625 products for the Baby Clothes dataset.

For each user, one feedback was randomly chosen for the test set, and the other feedbacks were chosen for the training set [15][56]. Furthermore, to demonstrate Deep-MINE's performance under cold start settings, cold start test sets including products with different sparsity levels were also extracted from the test set, which will be elaborated in Section 4.3.3.

4.2 Evaluation metrics and baseline models

Aligned with the prior literature, the Area Under the ROC Curve (AUC) [15][44] and Hit Ratio [58] were chosen as two metrics for performance evaluation. The AUC is defined as

$AUC = \frac{1}{|I|} \sum_i \frac{1}{|J|} \sum_{(i,j,j') \in S} \delta(x_{ij} > x_{ij'})$, and $\delta(\cdot)$ is an indicator function that equals 1, if $x_{ij} > x_{ij'}$ is true; otherwise it is 0. The AUC measures the ratio of correctly predicted product pairs to the total product pairs for all consumers. S is defined in the formulation in Section 3.1. The hit ratio is also widely used in recommender system evaluation [57]. The hit ratio is measured as the percentage of users who have at least one correctly recommended product in the top-K recommendation list. A higher hit ratio reflects a higher recommendation accuracy. In the following experiments, different K values were tested to prove the robustness of the proposed model.

To show the superior performance of Deep-MINE, the following baseline models were chosen for comparison.

- (1) BPRMF: The pairwise Bayesian Personalized Ranking model proposed in [44], which is a state-of-the-art ranking based model only utilizing implicit feedback data.
- (2) CDL: The Collaborative Deep Learning model proposed in [53] processes description content with a stacked denoising auto-encoder based on the probabilistic matrix factorization framework.
- (3) VBPR: The Visual Bayesian Personalized Ranking model proposed in [15], i.e., based on BPRMF, utilizes image features from pre-trained image classification model.
- (4) CKE: The Collaborative Knowledge Base Embedding model, i.e., an extension of CDL proposed in [56], incorporates structural, textual and visual content. As no structural information is available in our context, descriptions and images are considered here in implementation.

A validation set sampled from the training set was used to find the optimal hyperparameters for the Deep-MINE model and all the baseline models above. The hyperparameter settings of the Deep-MINE model are listed in Table 2. Based on the previous literature [56], the node numbers of the latent layer for images, descriptions and reviews were set the same, and experiments under different parameter settings were also conducted in Section 4.3.5 to justify the model's robustness.

Table 2. Hyperparameters Settings

Deep-MINE Model	Hyperparameters Setting
-----------------	-------------------------

Image auto-encoder	$N_{m1}=64, N_{m2}=64, N_{m3}=100, \lambda_m = \frac{1}{\#img_dim}$
Description auto-encoder	$N_{d1}=400, N_{d2}=100, \lambda_d = \frac{1}{\#des_dim}$
Review auto-encoder	$N_{r1}=400, N_{r2}=100, \lambda_r = \frac{1}{\#rev_dim}$
Regularization and Variance Parameters	$\lambda_\theta = 0.1, \lambda_\beta = 0.001, \lambda_{W_{fu}} = 0.001,$ $\lambda_w = \frac{1}{\#W_dim}, \lambda_q = \frac{1}{\#Q_dim}, \lambda_n = \frac{1}{\#N_dim}, \lambda_b = \lambda_t = \lambda_c = 0$

(Note: N_{m1} , N_{m2} , and N_{m3} refer to the number of hidden units for layer 1, 2, and 3 in the image auto-encoder, respectively; img_dim , des_dim , and rev_dim refer to the dimensions of image, description and review input, respectively; $\#W_dim$, $\#Q_dim$, and $\#N_dim$ refer to the dimensions of corresponding weight matrix W , Q , and N , respectively.)

4.3 Experiment results

4.3.1 Performance comparison of Deep-MINE and baseline models

Deep-MINE and baseline models were evaluated under different settings in this subsection. All the models were trained using the same strategy as introduced above. To make a fair comparison, the factor numbers of hidden information and integrated content were kept the same for Deep-MINE and baseline models. The AUC results are shown in Fig. 5. For the Women's Dresses dataset, the AUC of Deep-MINE was no less than 0.85, meaning that more than 85% of the pairwise rankings were predicted accurately. With total factor number increasing from 50 to 200, the Deep-MINE model consistently performed better than baseline models. BPRMF performed the worst as it utilized only the feedback data without considering content information. CKE had the second-best performance because it utilized more information compared with VBPR and CDL. Not surprisingly, VBPR had a slightly better performance than CDL as image features are more informative than descriptions for products, such as clothes.

To examine the robustness of Deep-MINE, experiments were also conducted on the Baby Clothes dataset (Fig. 5(b)). Deep-MINE still had a better performance than all the baselines. A notable difference was that VBPR also achieved a good performance with a slightly lower AUC than Deep-MINE (0.8025 vs 0.8030). On this dataset, as the visual input was a 4096-dimension feature vector instead of raw images, the design of image convolutional auto-encoder in the Deep-MINE model was deprecated. Still, the Deep-MINE performance was satisfactory, using the mechanism of multi-view information integration. However, the high-dimensional visual feature may have a

dominant effect over the other two views, which may explain the relatively poor performance of CDL and CKE. Consistently, BPRMF performed the worst in all the baselines. As the model performance was not sensitive to the factor number on both datasets, the factor number was fixed at 100 in the following experiments.

In addition to the AUC, the hit ratio is another metric that measures the ranking-based recommendation accuracy. When K varied from 50 to 200, the hit ratio was plotted in Fig. 6. On the Women's Dresses dataset, Deep-MINE beaten all the other baseline models and achieved a performance similar to that of CKE. On the Baby Clothes dataset, Deep-MINE performed the best across all K levels.

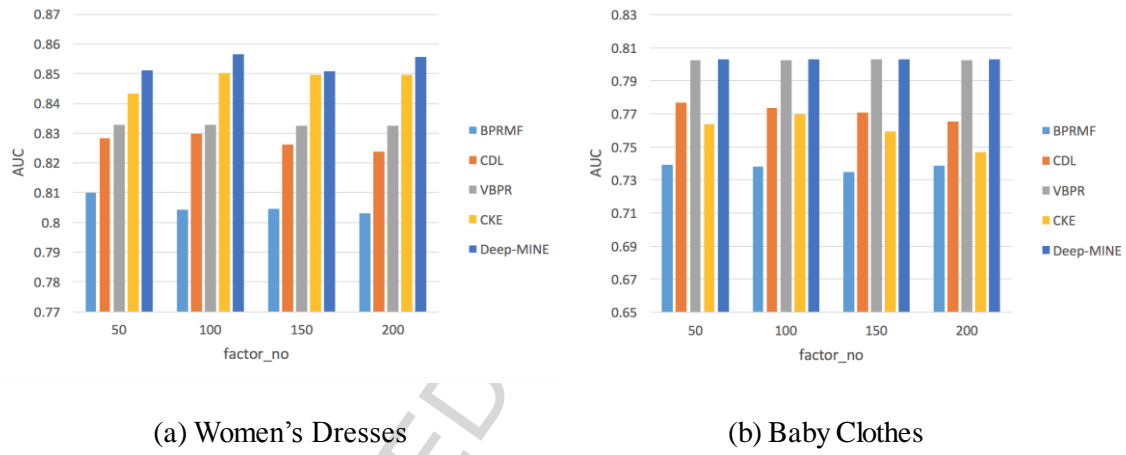


Fig. 5. AUC comparison of our model and baseline models.

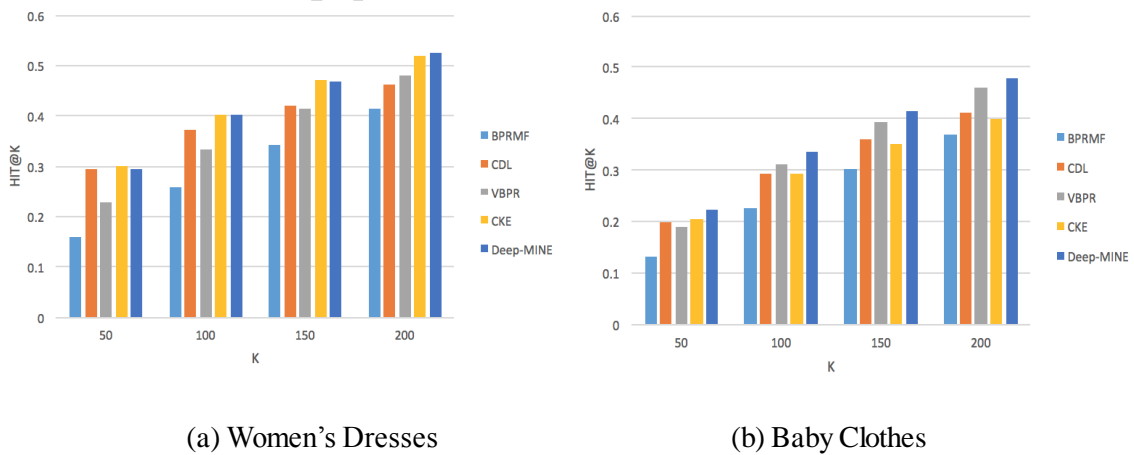


Fig. 6. Hit ratio comparison of our model and baseline models.

Through the above results, Deep-MINE revealed its effectiveness for integrating information in

comparison with other content-aware recommendation models. In addition, Deep-MINE also showed a significant improvement over BPRMF, which signifies the considerable advantages of incorporating content information into recommendation models.

4.3.2 Effect of multi-view information integration

As mentioned previously, Deep-MINE can organically integrate multi-view information to enhance recommendations. Therefore, Deep-MINE (i.e., the entire model) was further examined with its degenerated forms (i.e., single models), where one view from only a single information source was considered, namely, Image-MINE, Description-MINE and Review-MINE. From Fig. 7 and Fig. 8, it is clear that the entire model performed better than single models on the AUC and hit ratio. Concretely, Review-MINE performed better than the other single models, which may signify that reviews contain more valuable information compared with information provided by e-retailers, such as images and descriptions. Image-MINE and Description-MINE had slightly different performances on the two datasets. On the Women's Dresses dataset, product images mattered more, while for the Baby Clothes dataset, descriptions mattered more, possibly because baby clothes were more functional than women's dresses, and therefore, descriptions contained more relevant and decisive information for consumers. From an overall point of view, the proposed Deep-MINE model showed its advantage by organically integrating multi-view information. Furthermore, all the Deep-MINE related models performed significantly better than the BPRMF model.

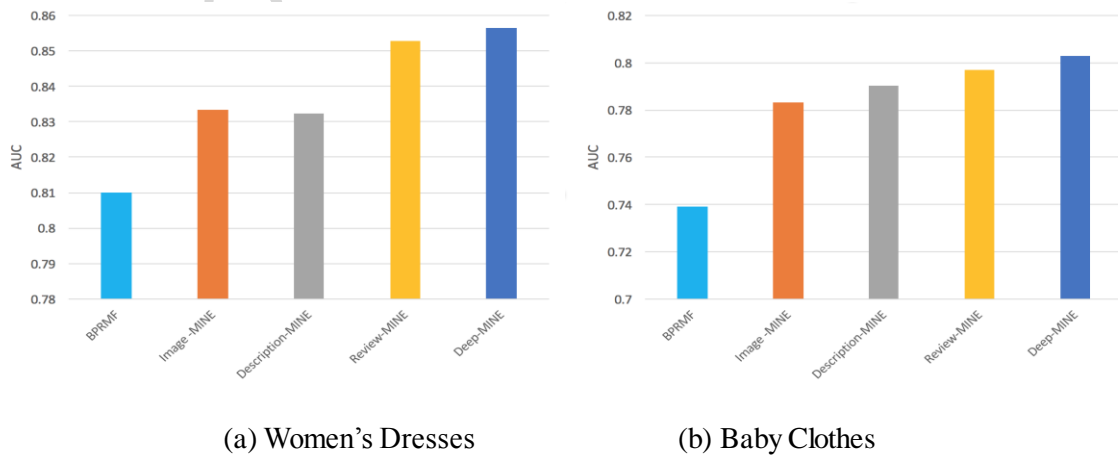


Fig.7. AUC comparison of the entire model and single models (Factor No = 100)

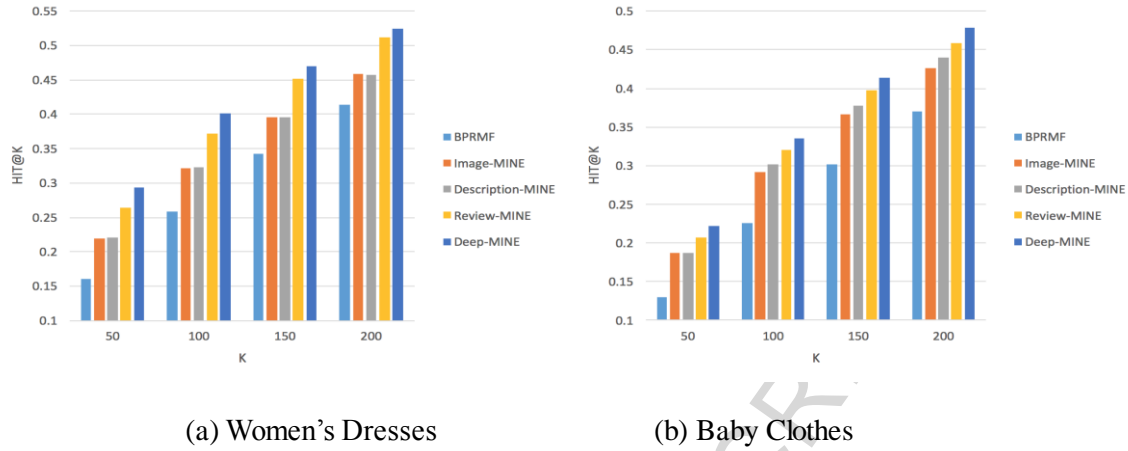


Fig. 8. Hit ratio comparison of the entire model and single models

4.3.3 Performance on cold start datasets

As Deep-MINE leverages multi-view content information to make recommendations, it is expected to suffer less in cold start product settings, i.e., products having no or few purchase feedbacks could hardly be recommended. To verify this, cold start test sets of different sparsity levels were extracted from the full test set. The “Sparsity Level=1” group refers to cold start products that have no historical feedback in the training set and only appear once in the test set. The “Sparsity Level=10” group refers to cold start products that have no more than 10 feedbacks in the dataset. All of the baseline models were tested, and results are listed in Table 3.

Table 3. AUC performance on test sets of different sparsity levels

Sparsity Level	1	2	3	4	5	6	7	8	9	10
BPRMF	0.3020	0.3368	0.3672	0.3950	0.4260	0.5732	0.6047	0.6275	0.6454	0.6587
CDL	0.3213	0.3601	0.3870	0.4307	0.4632	0.6245	0.6543	0.6751	0.6924	0.7046
VBPR	0.1441	0.2276	0.3066	0.3650	0.4148	0.6245	0.6520	0.6721	0.6910	0.7032
CKE	0.3843	0.4214	0.4484	0.4709	0.4940	0.6193	0.6484	0.6693	0.6864	0.6966
Deep-MINE	0.5001	0.5014	0.5377	0.5623	0.5799	0.6887	0.7085	0.7252	0.7390	0.7484

The results further demonstrate that through a better exploitation of visual and textual content, Deep-MINE outperformed all the baseline models with remarkable advantages. In particular, in the “Sparsity Level=1” group, the AUC of Deep-MINE surpassed that of the best baseline model (i.e., CKE) by 30.13% (0.5001 vs 0.3843). As feedbacks for each product increased, the recommendation performances improved for all models. CKE performed the best in the baseline models as it takes account of various content as well. As expected, BPRMF performed the worst as it only considers feedback data.

4.3.4 Effect of incorporating cognitive styles

To further examine the impact of incorporating the heterogeneity in user cognitive styles and demonstrate the reliability of the cognitive style values obtained by Deep-MINE, we proposed a series of initial cognitive indexes as benchmark indexes and compared their recommendation performances with the Deep-MINE model. Higher recommendation performance could imply, to some extent, that the corresponding cognitive index configuration reflects a more accurate representation of the user’s cognition characteristics.

The benchmark indexes include: (1) No cognitive styles. This configuration represents that cognitive style information is not considered in the model and recommendation. (2) Uniform cognitive styles. Uniform cognitive styles assume the cognitive weights of all consumers on the three pieces of information, i.e., descriptions, reviews and images, are the same, which is $[1/3, 1/3, 1/3]$. This index assumes that not only the cognitive styles of all consumers are homogenous, but also the consumers’ inclinations to the three different forms of information representations are indifferent. (3) Ordered cognitive styles. This index assumes all consumers have the same priority order on different information formats, i.e., with weight $[3/6]$ being high priority, $[2/6]$ being medium priority and $[1/6]$ being low priority, thus generating 6 combinations of cognitive style indexes, e.g., [high-image, medium-description, low-review], [high-description, medium-review, low-image], etc. (4) Random cognitive styles. A randomization could ensure that each consumer has a differentiated cognition vector. This is a general treatment to represent consumer’s cognition heterogeneity without further information. (5) Average cognitive styles. In this configuration, the cognitive style of each consumer

is derived using our proposed model. Then the average cognitive style values across all consumers are calculated and used as the unified cognitive style.

The comparison results on the hit ratio and AUC are shown in Fig. 9 (a) and (b), and clearly the proposed Deep-MINE showed the best performance. Further findings can also be derived. First, all the models equipped with cognitive styles significantly outperformed the model without cognitive styles, further emphasizing the power of integrating cognitive styles into personalized recommendation. Second, different cognitive style distributions on three information representations did significantly impact the recommendation performance, which demonstrates the importance of detecting appropriate cognitive styles of consumers. Third, the Deep-MINE model (which treats consumer's cognitive style in an individualized manner) outperformed other models only considering unified cognitive style across all consumers, showing the strength of cognitive style personalization. It is also worth mentioning that, though configuration (5) of Average Cognitive Styles to a large extent incorporates the learnt cognitive styles of all consumers, the final performance was weakened due to the average treatment compared with the Deep-MINE model.

To visualize the cognition values, an extracted sample of ten consumers' cognition value distributions from the Women's Dresses dataset is shown in Fig. 9(c), in which we observed that consumers 1 and 8 cared much more about images and reviews than descriptions; consumers 2, 3, 5 and 9 were more inclined to descriptions; consumer 6 valued reviews more; consumer 7 valued images more; and consumers 4 and 10 paid roughly equal attention to all three views. Such observations further confirm the prevalence of cognition heterogeneity, and highlight the significance of treating multiple information sources differently for different types of consumers in recommendations. For instance, from the perspective of online shopping platform, platform managers could consider a personalized webpage layout design that is consistent with the consumer's individualized cognitive style to provide a better shopping experience, which is also supported by Engin and Vetschera [11].

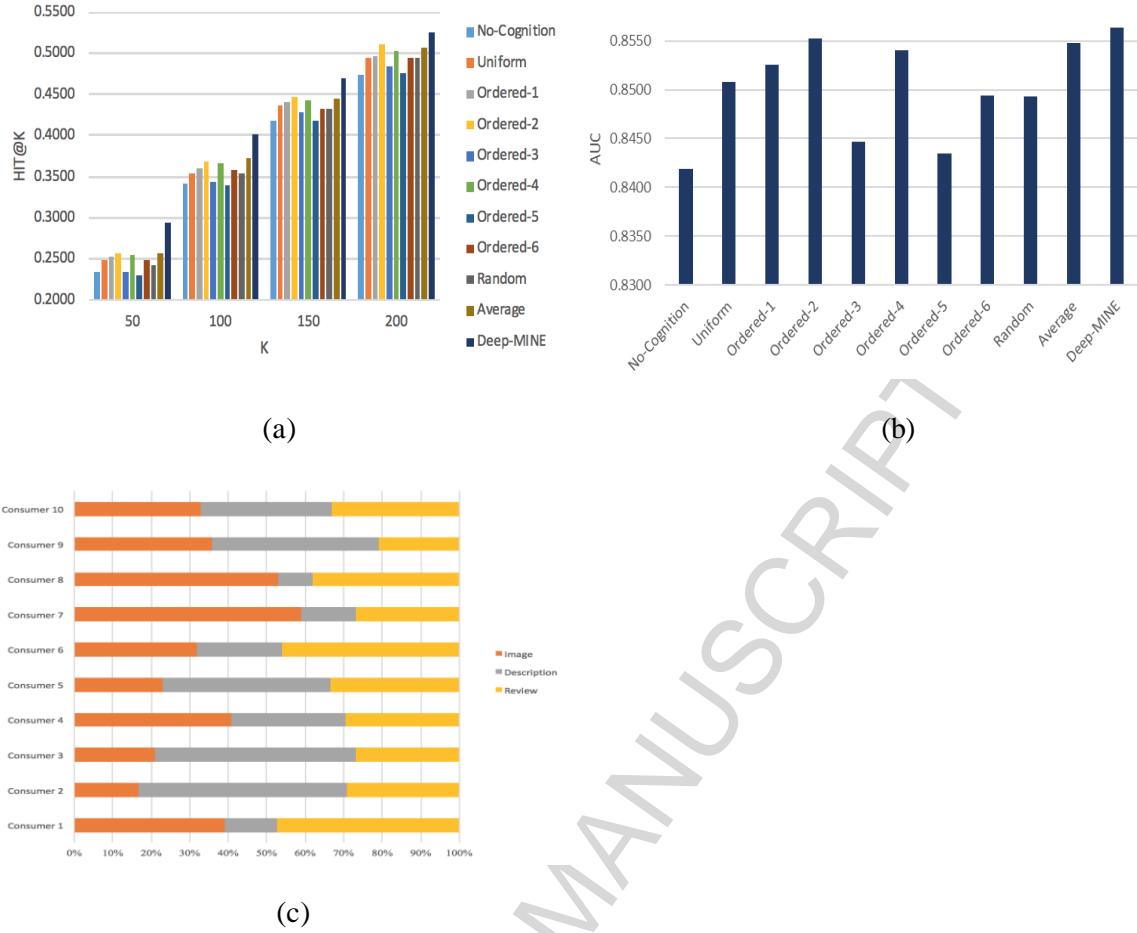


Fig. 9. The effect of incorporating cognitive styles. (a) Hit ratio performance; (b) AUC performance; (c) 10 random consumers' cognition style distribution

4.3.5 Sensitivity Analysis

As mentioned in Section 4.2, the node numbers of the latent layer for three auto-encoders were all set to 100. To further prove the robustness, we also conducted experiments by varying the latent node number for the image auto-encoder (Table 4). It was observed that the performance remained stable across different numbers, and more image nodes did not necessarily lead to a better performance, possibly because the learning process can ensure an effective representation and integration of all information through back propagation to minimize the objective function, regardless of the initial hyperparameter settings.

Table 4. Model performance with different node numbers for images

Image Auto-encoder	AUC	HIT@50	HIT@100	HIT@150	HIT@200
50-Node	0.8564	0.2816	0.3986	0.4742	0.5345
100-Node	0.8564	0.2936	0.4009	0.4695	0.5250
200-Node	0.8554	0.2826	0.3901	0.4631	0.5198
300-Node	0.8526	0.2821	0.3941	0.4685	0.5227

A 6-layer convolutional auto-encoder was designed in this study to represent information on product images, which is consistent with [56]. Some trial experiments were also conducted to help determine the number of layers. Specifically, a subset of product images was selected from the training set, and networks with different number of layers were trained to achieve their local optimum. To compare their relative performance, we measured their loss function values, (i.e., as defined in Section 3.2.1 denoted by $L1$, which consists of reconstruction error terms and regularization terms for the weight matrix). It is obvious from Table 5 that the 6-layer structure had a better performance than the 4-layer and 8-layer networks.

Table 5. Loss function values with different layers of image auto-encoder

Image Auto-encoder	4-Layer	6-Layer	8-Layer
Loss Function Value	0.0247	0.0137	0.0567

4.3.6 Visualization of the recommendation results

To better illustrate the recommendation performance of Deep-MINE, four consumers in the Women's Dresses data were randomly selected with the top-5 recommendation image results generated by Deep-MINE, BPRMF, CDL, VBPR and CKE, respectively (see Fig. 10). The first row is the dresses that the consumers previously purchased, reflecting the consumers' historical tastes. The other five rows are the top-5 dresses recommended by Deep-MINE, BPRMF, CDL, VBPR and CKE, respectively. It can be intuitively observed that Deep-MINE had more recommendation variety compared with the baseline models. Concretely, BPRMF and VBPR tended to recommend the most popular products without much personalization for different consumers, i.e., 3–4 popular dresses could be repeatedly found in the recommendation results for the four consumers. In addition,

Deep-MINE recommended more relevant dresses based on the tastes reflected in the consumers' historical purchases, such as color (e.g., dark or colorful), size (e.g., long or short) and style (e.g., casual or formal). Although CKE and CDL showed some variety, they were not very consistent in taste with consumers' previous purchases.



Fig. 10. Visualization of the top-5 recommendations by Deep-MINE and baseline models.

5 Conclusion and future work

This study proposed a personalized recommendation model with multi-view information integration, i.e., Deep-MINE, which organically and comprehensively utilizes multiple sources of product content and considers users' heterogeneity in cognitive styles. A unified deep neural network was designed as an end-to-end model composed of three main parts: multi-view information representation, cognition treatment, and information integration, by which preferable recommendation performances can be achieved. Extensive data experiments revealed the better of Deep-MINE in comparison to the baseline models. This study also shed light on the potential of a data-driven view of cognitive style.

Future work could be extended in the following directions. First, this study only considered a single kind of implicit feedback, namely, consumer purchase behavior. Some other feedback, such as product

returns, consumer browsing and clicking behavior, provides more detailed information about consumer preferences and thus could be further exploited in the design of recommender systems. For instance, the information representation part in the Deep-MINE model could be enriched by integrating such knowledge. Furthermore, this study incorporated the heterogeneity of cognitive style into a recommender system and proposed an integrated deep learning framework to solve the problem. Future research may consider representing user's cognitive styles from the perspective of other dimensions in addition to the Verbal-Imagery dimension considered in this study, to enrich the measurement of cognitive styles.

Acknowledgements

The work was partly supported by the National Natural Science Foundation of China (71490724/71772101) 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] C. Allinson, J. Hayes, The Cognitive Style Index: a measure of intuition-analysis for organizational research, *Journal of Management Studies* 33 (1) (1996) 119-135.
- [3] R. Bendall, A. Galpin, L. Marrow, S. Cassidy, Cognitive style: time to experiment, *Frontiers in Psychology* 7 (2016) 1786.
- [4] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (2013) 1798–1828.
- [5] S. Boutemedjet, D. Ziou, A graphical model for context-aware visual content recommendation, *IEEE Transactions on Multimedia* 10 (2008) 52–62.
- [6] I. Chakraborty, P. Hu, D. Cui, Examining the effects of cognitive style in individuals' technology use decision making, *Decision Support Systems* 45 (2008) 228-241.

- [7] M. Chen, Improving website structure through reducing information overload, *Decision Support Systems* 110 (2018) 84–94.
- [8] C. Cheng, H. Yang, M.R. Lyu, I. King, Where you like to go next: successive point-of-interest recommendation. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 2605–2611.
- [9] T.L. Childers, M.J. Houston, Conditions for a picture-superiority effect on consumer memory, *Journal of Consumer Research* 11 (1984) 643–654.
- [10] F. Coffield, D. Moseley, E. Hall, K. Ecclestone, *Learning styles and pedagogy in post-16 learning: a systematic and critical review*, London: Learning & Skills Research Centre (2004).
- [11] A. Engin, R. Vetschera, Information representation in decision making: the impact of cognitive style and depletion effects, *Decision Support Systems* 103 (2017) 94–103.
- [12] M. De Gemmis, P. Lops, C. Musto, F. Narducci, G. Semeraro, *Recommender systems handbook*, 2nd ed, Springer, 2015.
- [13] J.R. Hauser, G.L. Urban, G. Liberali, M. Braun, Website morphing, *Marketing Science* 28(2) (2009) 202–223.
- [14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [15] R. He, J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016, pp. 144–150.
- [16] R. He, J. McAuley, Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016b, pp. 507–517.
- [17] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks. *Proceedings of the Fourth International Conference on Learning Representations (ICLR)*, 2015.
- [18] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, *Neural Computing* 18 (2006) 1527–1554.

- [19] Y. Hu, C. Volinsky, Y. Koren, Collaborative filtering for implicit feedback datasets. Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM), 2008, pp. 263–272.
- [20] Z. Jiang, I. Benbasat, Virtual product experience: effects of visual and functional control of products on perceived diagnosticity and flow in electronic shopping, *Journal of Management Information Systems* 21 (2004) 111–147.
- [21] X. Jin, J. Luo, J. Yu, G. Wang, D. Joshi, J. Han, Reinforced similarity integration in image-rich information networks, *IEEE Transactions on Knowledge & Data Engineering* 25 (2013) 448–460.
- [22] Q. Jones, G. Ravid, S. Rafaeli, Information overload and the message dynamics of online interaction spaces: a theoretical model and empirical exploration, *Information Systems Research* 15 (2004) 194–211.
- [23] J. Kisielius, B. Sternthal, Detecting and explaining vividness effects in attitudinal judgments, *Journal of Marketing Research* 21 (1984) 54–64.
- [24] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, *Computer* 8 (2009) 30–37.
- [25] M. Kozhevnikov, Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style, *Psychological Bulletin* 133(3) (2007) 464–481.
- [26] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.
- [27] Y.A. LeCun, Y. Bengio, G.E. Hinton, Deep learning, *Nature*. 521 (2015) 436–444.
- [28] C. Lei, D. Liu, W. Li, Z. Zha, H. Li, Comparative deep learning of hybrid representations for image recommendations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2545–2553.
- [29] N.H. Leonard, R.W. Scholl, K.B. Kowalski, Information processing style and decision making, *Journal of Organizational Behavior* (1999) 407–420.
- [30] K.H. Lim, I. Benbasat, The effect of multimedia on perceived equivocality and perceived usefulness of information systems, *MIS Quarterly*. 24 (2000) 449–471.
- [31] 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.
- [32] 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.
- [33] 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.
- [34] Q. Liu, S. Wu, L. Wang, DeepStyle: learning user preferences for visual recommendation. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 841–844.
- [35] S. Lu, L. Xiao, M. Ding, A video-based automated recommender (VAR) system for garments, *Marketing Science* 35 (2016) 484–510.
- [36] 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.
- [37] J. Masci, U. Meier, D. Cireşan, J. Schmidhuber, Stacked convolutional auto-encoders for hierarchical feature extraction. *International Conference on Artificial Neural Networks*, 2011, pp. 52–59.
- [38] J. McAuley, J. Leskovec, Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems (RecSys)*, 2013, pp. 165–172.
- [39] J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel, Image-based recommendations on styles and substitutes. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015, pp. 43–52.
- [40] S. Messick, The nature of cognitive styles: problems and promise in educational practice, *Educational psychologist* 19(2) (1984) 59-74.
- [41] Y. Ouyang, W. Liu, W. Rong, Z. Xiong, O. Yuanxin, L. Wenqi, R. Wenge, X. Zhang, Autoencoder-based collaborative filtering. *International Conference on Neural Information Processing*, 2014, pp. 284–291.
- [42] W. Pan, L. Chen, GBPR: group preference based Bayesian personalized ranking for one-class collaborative filtering. *Proceedings of the Twenty-Third International Joint Conference on*

- Artificial Intelligence (IJCAI), 2013, pp. 2691–2697.
- [43] 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.
- [44] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback. *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 2009, pp. 452–461.
- [45] R. Riding, I. Cheema, Cognitive styles—an overview and integration, *Educational psychology* 11(3-4) (1991) 193-215.
- [46] R. Riding, S. Rayner, Cognitive styles and learning strategies: understanding style differences in learning and behavior, David Fulton Publishers, 2013.
- [47] R. Salakhutdinov, A. Mnih, G. Hinton, Restricted Boltzmann machines for collaborative filtering. *Proceedings of the 24th international conference on Machine learning (ICML)*, 2007, pp. 791–798.
- [48] M. Siering, A.V. Deokar, C. Janze, Disentangling consumer recommendations: explaining and predicting airline recommendations based on online reviews, *Decision Support Systems* 107 (2018) 52-63.
- [49] J.Z. Sojka, J.L. Giese, The Influence of personality traits on the processing of visual and verbal information, *Marketing Letters* 12 (2001) 91–106.
- [50] J.Z. Sojka, J.L. Giese, Communicating through pictures and words: understanding the role of affect and cognition in processing visual and verbal information, *Psychology & Marketing* 23(12) (2006) 995–1014.
- [51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, C. Hill, A.S. Arora, Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 1–9.
- [52] C. Wang, D.M. Blei, Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 2011, pp. 448-456.
- [53] H. Wang, N. Wang, D.-Y. Yeung, Collaborative deep learning for recommender systems.

- Proceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), 2015, pp. 1235–1244.
- [54] M. Wang, X. Li, P.Y. Chau, The impact of photo aesthetics on online consumer shopping behavior: an image-processing-enabled empirical study. Proceedings of 2016 International Conference on Information Systems (ICIS), 2016.
- [55] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, H. Liu, What your images reveal: exploiting visual contents for point-of-interest recommendation. Proceedings of the 26th International Conference on World Wide Web (WWW), 2017, pp. 391–400.
- [56] F. Zhang, N.J. Yuan, D. Lian, X. Xie, W. Ma, Collaborative knowledge base embedding for recommender systems. Proceedings of the 22th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), 2016, pp. 353–362.
- [57] S. Zhang, L. Yao, Deep learning based recommender system: a survey and new perspectives, ACM Computing Surveys 1(1) (2018).
- [58] Y. Zhang, Q. Ai, X. Chen, W.B. Croft, Joint representation learning for top-N recommendation with heterogeneous information sources. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM), 2017, pp. 1449–1458.

Bibliographical Note

YueGuan is currently pursuing her PhD degree at the School of Economics and Management, Tsinghua University, Beijing, China. Her research interests include deep learning, business analytics, and online recommendation.

Qiang Wei is an associate professor in the Department of Management Science and Engineering, School of Economics and Management, Tsinghua University, Beijing, China. His current research interests include deep learning, information management, business analytics, and business intelligence.

Guoqing Chen received his PhD from the Catholic University of Leuven (K.U. Leuven, Belgium) and now is Professor of Information Systems at the School of Economics and Management, Tsinghua University, Beijing, China. His research interests include information systems management, business analytics and decision support systems.

Highlights

- 1) An end-to-end deep-learning based recommendation model Deep-MINE is proposed by organically integrating multi-view information (i.e., visual content, textual content, etc.)
- 2) Both user and content heterogeneity are well addressed and taken into account in design of the model framework.
- 3) Stacked auto-encoder networks are developed to map heterogeneous information into a unified latent space.
- 4) Extensive experiments and visual demonstrations prove the outperformance of the proposed model.