Information Sciences 519 (2020) 306–316

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A novel hybrid deep recommendation system to differentiate user's preference and item's attractiveness

Xiaofeng Zhang[a,*], Huijie Liu[a], Xiaoyun Chen[b], Jingbin Zhong[a], Di Wang[c]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[b] Faculty of Business Administration, University of Macau, China
[c] Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 7 August 2018
Revised 26 January 2020
Accepted 28 January 2020
Available online 28 January 2020

Keywords:
Probabilistic matrix factorization
Deep learning
Recommendation systems

ABSTRACT

With the fast development of online E-commerce Websites and mobile applications, users' auxiliary information as well as products' textual information can be easily collected to form a vast amount of training data. Therefore, research efforts are urgently needed to make customized recommendations using such large but sparse data. Deep recommendation model is a natural choice for this research issue. However, most existing approaches try to investigate either user's auxiliary information such as age and zipcode, or item's textual information such as product descriptions, reviews or comments. Therefore, it is desired to see whether user's auxiliary information and item's textual information could be modeled simultaneously. This paper proposes a novel approach which is essentially a hybrid probabilistic matrix factorization model. Particularly, it has two sub components. One component tries to predict user's rating scores by capturing user's personal preferences extracted from auxiliary information. Another component tries to model item's textual attractiveness to different users via a proposed attention based convolutional neural network. We then propose a global objective function and optimize these two sub components under a unified framework. Extensive experiments are performed on five real-world datasets, i.e., ML-100K, ML-1M, ML-10M, AIV and Amazon sub dataset. The promising experimental results have demonstrated the superiority of our proposed approach when compared with both baseline models and state-of-the-art deep recommendation approaches, i.e., PMF, CDL, CTR, ConvMF, ConvMF+ and D-Attn with respect to RMSE criterion.

1. Introduction

Recently, with the fast development of various online E-commerce Websites and mobile applications, users can easily share their reviews or comments which can be collected to form a vast amount of textual data reflecting users' subjective opinions. Generally, users would like to seek various reviews/comments for comparison before they purchase some products. On the other hand, users may have their own inherent preferences which might affect purchase decisions. Apparently, users' potential preferences could be partially discovered from their personal information such as age, education level and zipcode. Such information is already released to public access when users register on E-commerce Websites. Accordingly, it has become an emerging research topic to make recommendation by utilizing either product's comments or user's auxiliary information [10,35].

* Corresponding author.
E-mail address: zhangxiaofeng@hit.edu.cn (X. Zhang).

https://doi.org/10.1016/j.ins.2020.01.044
0020-0255/

Conventionally, most existing recommendation systems predict users' rating scores using either item-based or user-based approaches. These approaches utilize the calculated similarity between items or users to predict scores of unrated items. However, the performance of these approaches is inevitably poor when the scored items are very few. In addition, it is hard to give a diversified recommendation. The recent proposed Probabilistic Matrix Factorization (PMF) [19] is good at coping with this situation. Unfortunately, the ordinary PMF does not consider the effect of users' auxiliary information or product's comments when predict the scores. Given a large amount of spare data, it is desired to propose deep recommendation approach which can achieve robust performance. Thus, deep learning based approaches are proposed in recent years. However, neither of them has modeled product's comments and user's auxiliary information simultaneously.

As aforementioned, it is a common practice that customers would prefer to read comments before they buy certain products. However, different users may have different subjective opinions after reading the same product comments. We assume that such emotional tendencies may be caused by different characteristics of users. These characteristics could be partially discovered from users' auxiliary information. For example, young people may like to watch "Action" movies, whereas female audience may like "Romantic" movies. Although this is not always the case, we can learn these potential preferences for item recommendations.

This paper is proposed for this research issue. Particularly, a hybrid PMF based framework is proposed which assumes that different user-specific latent features can well reflect users' preferences. In the meanwhile, it models how different item-specific latent features can well attract different types of users. To extract user-specific latent features, a stacked denoising autoencoder is adopted as one sub component and is optimized globally. For the extraction of item-specific latent features, it is not appropriate to directly adopt all terms appeared in the comments. A convolutional neural network could be used for the extraction of semantic meaning of a document. Without loss of generality, we assume that users may be interested in different textual content contained in product comments. This means that users may only pay attention to one or several terms in the comments. In this case, we carefully design an attention based CNN to represent item-specific latent features. The two sub components are then modeled under a unified framework which globally optimizes the proposed loss function and iteratively updates the parameter set.

The main contributions of this paper can be summarized as follows:

- In this paper, a novel hybrid PMF model is proposed called DUPIA which can simultaneously differentiate user's preference and item's attractiveness. To the best of our knowledge, this is the first attempt to predict rating scores by discovering users' inherent preferences and their emotional tendencies simultaneously.
- Two sub deep learning components of DUPIA are designed and a global objective function is proposed to optimize two sub components under a unified learning framework. Particularly, the attention based CNN component can best capture the discriminative item-specific latent features for item recommendation.
- Extensive experiments are evaluated on several real-world datasets. The proposed DUPIA approach is superior to the baseline PMF, CDL [29], and the state-of-the-art algorithms CTR [26], ConvMF, ConvMF+ [13] and D-Attn [20] with respect to RMSE. Furthermore, the extracted important terms from the attention based CNN, a sub component of DUPIA, has demonstrated that they can well differentiate item's attractiveness to different users, and this further verifies the effectiveness of the proposed approach.

The remainder of this paper is organized as follows. Section 2 introduces related work on deep recommendation systems and Section 3 formulates the raised research issue and then proposes a novel unified framework for this issue. Extensive experiments are performed in Section 4 and we concludes the paper in Section 5.

2. Related work

In this section, we first briefly review conventional techniques for recommendation systems as well as recent deep learning based approaches.

2.1. Conventional recommendation systems

At the early stage, recommendation algorithms can be classified into item based approach and user based approach [1,4]. It is well known that these approaches are proposed based on the calculated similarities among either users or items. Then, the scores of unrated items are predicted based on the rating scores of similar items or users [8,9]. Various algorithms have been proposed in the literature. Later, various collaborative filtering (CF) [21,32] based approaches have been proposed. In these approaches, the core component is to design a mechanism to predict rating scores based on the similar groups of users or items. Among them, one of the most successful models, probabilistic matrix factorization (PMF) [19], is proposed which tries to find low rank representation to represent the relationship between a large user matrix and item matrix. These low rank representations can well interpret users' preferences on some items.

To model users' auxiliary information, several PMF based approaches have been proposed. For instance, several hybrid collaborative filtering based approaches are proposed to take side information [14] into consideration in the process of matrix factorization, such as Collective Matrix Factorization (CMF) [22] and [16]. However, the side information is treated by these approaches only as regularization term and is not for providing hints to guess users' preferences. Wang et al. proposed collaborative topic regression (CTR [26]) approach which integrates probabilistic matrix factorization (PMF) with

topic model (Latent Dirichlet Allocation) [3]. By considering items' textual information, CTR can learn features from the side information and then makes recommendations. In some emerging research areas such as mobile crowdsensing, there also exist prediction-based user recommendation task [27,28]. For example, in [28], the proposed Prediction-based User Recruitment (PURE) approach can separate users into two different groups which might be interested in different price plans. Similarly in [27], users of mobile crowdsensing are selected to satisfy both temporal and spatial constraints and whether the selected users prefer to accomplish sensing data also depends on data properties. Therefore, a triple-layer structure model is proposed which considers temporal, spatial and data information. A greedy algorithm is proposed to resolve this optimization problem.

2.2. Deep recommendation systems

Encouraged by the big success of deep learning in the research areas like computer vision and natural language processing [2,5,12,34], various deep learning based approaches are proposed for recommendation systems. At the very beginning, researchers try to design deep learning framework to extract latent features from side information. For example, Oord et al. [17] uses deep convolutional neural networks to generate latent factors for songs recommendation using their audio data without any rated data. Wang and Wang [30] proposes a combined model using both deep belief neural networks and a probabilistic graphical model to learn audio features. The recommendation is given in a customized manner. However, both of these aforementioned approaches are deterministic ones and have not considered the existence of noise and thus are susceptible to noisy data. Furthermore, inspired by the tightly coupled model CTR and stacked denoising autoencoder (SDAE) [25], a collaborative deep learning model is proposed called CDL [29] which learns latent factors from item descriptions. It is an efficient model but needs extra efforts to fine tune a set of hyper-parameters. The advantage of this approach lies in that it ignores the sequence order of the words due to the bag-of-words representation format. Kim etc. [13] utilizes a CNN for document processing and combine CNN with PMF, called ConvMF. The ConvMF can effectively capture contextual information and consequently enhances the prediction accuracy. Moreover, there are some other deep learning based attempts [7,13,24,36]. However, these aforementioned models merely extract latent features out of items' auxiliary information.

To achieve a better model performance, a number of hybrid models are proposed [6,20,31,33]. It is believed that the hybrid model is generally able to take advantage of each single model and thus can achieve superior model performance. In [31], the authors propose a hybrid model with two deep neural networks as its sub components to make a personalized tag-aware recommendation. This model utilizes both user and item profiles, e.g. tags, and converts them into ordinary latent feature space. The model performance is significantly improved by using the auto-encoder based approach. However, the tag data might be its limitation. Dong et al. [6] also proposes a hybrid deep learning model to utilize item and user side information simultaneously to alleviate the sparsity issue generated in user-item rating matrix. Additional Stacked Denoising Autoencoder(aSDAE) is adopted to convert side information into latent features. The side information used in this model is structural information. Practically, unstructured information is a common. Seo et al. [20] proposes local and global attention mechanism based CNN model to capture user preferences and item attractiveness from its reviews. However, the same model structure of user and item may contribute little to the performance of the hybrid model. To summarize, most existing approaches seldom use users' preferences and items' attractiveness simultaneously. This paper attempts to resolve this challenging issue. Inspired by Seo et al. [20], we first design our own attention based CNN model to capture features that are more likely to attract users. Different from existing approaches, our work carefully design a unified framework to associate users' side information with the weighted features of items' comments which is then optimized in a global manner.

3. The proposed DUPIA approach

In this section, we first detail the proposed novel approach termed Differentiating Users' Preferences and Items' Attractiveness (DUPIA) for recommendation systems. As aforementioned, the purpose of the proposed DUPIA is to predict users' scores by estimating diverse users' preferences as well as their interests in different perspectives of products. Conventionally, user's preferences are discovered from their rating scores which might be problematic if the rating scores are very sparse. Therefore, it is desired to extract user's preferences merely from their static attributes or auxiliary information such as age and education level. Similar to aSDAE, we also extend the stacked denoising autoencoder model (SDAE) to capture users' preferences from their auxiliary information. Meanwhile, we argue that different users might be attracted by different part of contextual information if they read the same product descriptions. In this case, we attempt to quantify product contextual attractiveness to different users through the attention mechanism. Particularly, a convolutional neural network (CNN) with a global attention layer is adopted to extract the core words out of each review of a product. The extracted core words are then modeled as the attractiveness of each product. The proposed hybrid model can thus model both users' preferences and products' attractiveness under a unified framework and the framework of the proposed approach is plotted in Fig. 1.

3.1. Problem formulation

Let R denote the rating score matrix, U and V respectively denote the user's latent factor and item's latent factor and we have $R = U * V$. S and X respectively denote users' rating scores and users' auxiliary information with noise, S' and X' respectively denote the reconstructed rating scores and users' auxiliary information, h_i denote the output of hidden layer.

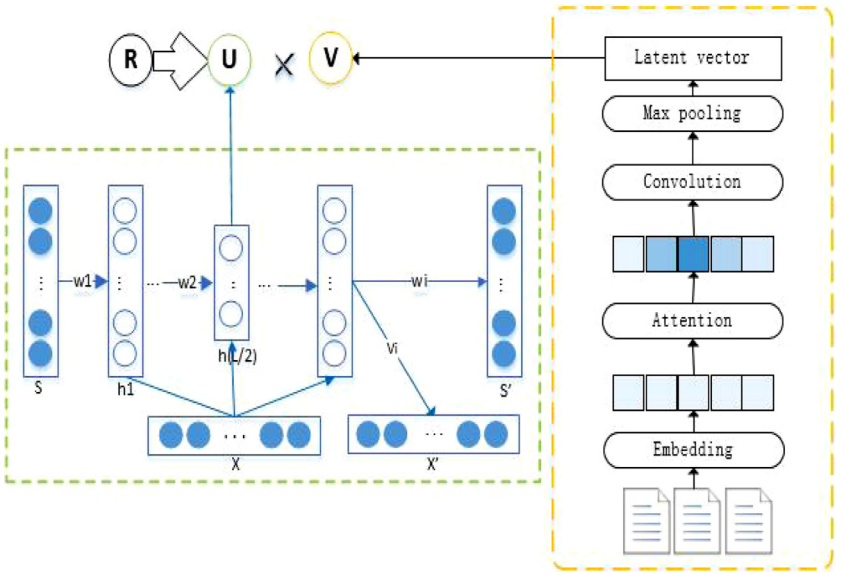

Fig. 1. The proposed DUPIA model.

Hidden layer is used to encode and decode both S and X, W_i and V_i denote the corresponding weights of each layer. To capture product's attractiveness, users' review on some items are modeled using D.

Essentially, the proposed DUPIA is a hybrid probabilistic matrix factorization (PMF) based approach which captures users' preferences and products' attractiveness simultaneously. Assume that we have N users and M items, the rating score matrix is denoted as $R \in R^{N \times M}$. In general, the proposed approach iteratively calculates user-specific and item-specific latent features. $U \in R^{D \times N}$, $V \in R^{D \times M}$ and their product($U^T V$) are accordingly used to approximate original rating matrix R. Let D denote the dimension of the latent matrix, the conditional probabilistic distribution over rating scores is defined as,

$$p(R|U, V, \sigma^2) = \prod_{i}^{N} \prod_{j}^{M} N\left(r_{ij}|u_i^T v_j, \sigma^2\right)^{I_{ij}}, \tag{1}$$

where $N(x|\mu, \sigma^2)$ is the probability density function of the Gaussian distribution and μ and σ^2 denote its mean and variance, and I_{ij} is the indicator function whether the ith user rates the jth item or not. In the following subsections, we first illustrate how to model users' preferences as well as products' attractiveness and then detail the optimization process.

3.2. Modeling users' preferences

The original additional stacked denoising autoencoder (aSDAE) is proposed to extract users' diverse preferences through their auxiliary information. As users' auxiliary information such as age and education level are generally well-structured information, aSDAE can learn robust representations by encoding and decoding the original users' attributes through hidden layers. As shown in left part of Fig. 1, aSDAE has two input variables, i.e., S and X. S' and X' are the reconstructed variables through the encoding and decoding layers. Let (h_1, h_L) represent hidden layers, and $(h_1, h_{\frac{L}{2}})$ denote encoder component. The last $\frac{L}{2}$ layer is decoder component. The encoder $g(\cdot)$ also has two inputs S and X, and it maps these inputs via a function $g(s, x)$. The decoder $f(\cdot)$ reconstructs the hidden representation via function $f(g(s, x))$. The original aSDAE minimizes the error between original data and reconstructed data. The layer $h_{\frac{L}{2}}$ can be considered as users' latent features. Both $g(\cdot)$ and $f(\cdot)$ choose the same activation function for encoding and decoding layers. Accordingly, the hidden representation h_i of each hidden layer $l \in \{1, \ldots, L-1\}$ can be computed as follows

$$h_l = g(W_l h_{l-1} + V_l X + b_l), \tag{2}$$

where $h_0 = S$, S and X are twisted data with random noises, and S_0, X_0 are the original data. The output variables S' and X' can then be calculated as

$$S' = f(W_L h_L + b_{s'}) \tag{3}$$

$$X' = f(V_L h_L + b_{x'}) \tag{4}$$

3.3. Modeling items' attractiveness

To capture items' attractiveness to different users, it is a natural choice to adopt CNN based approach. As aforementioned, we assume that different users may be interested in different contextual information if they read the same piece of comment or review. In this case, CNN is adopted to assign different weights to different terms contained in a document (review). As is known, the importance of the core words is usually overwhelmed by the large amount of ordinary terms. To alleviate this issue, attention mechanism is introduced and an attention based CNN is proposed to cope with this issue. We expect that a small number of irrelevant terms could be extracted out of each document which will play an important role in attracting different users. Particularly, the proposed attentional based CNN model contains five layers, i.e., embedding layer, attention layer, convolution layer, max pooling layer and output layer. We will detail these components in the following paragraphs.

In the embedding layer, we adopt the state-of-the-art word vector to represent each document. First, we construct a vocabulary set according to the document frequency of each term. Hence, the term with a higher word frequency will be assigned with a smaller id in the vocabulary set. Suppose the size of the vocabulary set is l, and the embedding dimension is d. Any document D_i can be treated as a word sequence with length l and D_i will be transferred to a numeric matrix $W \in R^{dl}$. Each word in this document can be represented as a word vector, denoted as $w_i \in R^d$. The numeric matrix W is initialized using a pre-trained word vector like in Glove [18].

The introduced attention layer is to learn the importance of a word within a fixed-width window. Suppose the window size is l, and the center word is t_i. Accordingly, the word sequence in this window is T_i which can be calculated as

$$T_i = (t_{i+\frac{-l+1}{2}}, t_{i+\frac{-l+3}{2}}, \ldots, t_i, \ldots, t_{i+\frac{l+-1}{2}}). \tag{5}$$

The importance score $s(i)$ of term t_i can be computed as follows

$$s(i) = g(T_i * W_{att} + b_{att}), \tag{6}$$

where W_{att} and b_{att} are respectively the weight matrix and the bias, and $g(\cdot)$ is the activation function. Consequently, the weight sequence, denoted as T^A, of the whole embedded word vectors can be calculated as

$$T_i^A = s(i) w_i, \tag{7}$$

where T_i^A is the ith element in this sequence and i falls in [1, l].

After acquiring the word importance calculated in the attention layer, we can further compute textual features using the convolution layer. Let C denote the feature, W_c is the shared weight and b_c is the bias. The kernel size is l_c, and C can be calculated as,

$$C(i) = f(T_i^A * W_c + b_c), \tag{8}$$

where $f(\cdot)$ is activation function. Let C_{out} denote the eventual output of textual features which is acquired after a pooling layer, which is given as

$$C_{out}(i) = MAX(C(i)). \tag{9}$$

Hereinafter, we denote this textual feature as $C_{out}(W, T_j)$.

3.4. The proposed DUPIA approach

By considering both users' preferences and items' attractiveness, the objective function of the proposed approach can be written in Eq. (1), where u and v refer to user-specific and item-specific latent features, respectively. Without loss of generality, these latent features are assumed to subject to Gaussian distribution, and the conditional distribution over user-specific latent features can be written as,

$$p\left(U|W, V, \sigma_u^2\right) = \prod_i^N \left(u_i | h_{\frac{l}{2}i}, \sigma_u^2 I\right), \tag{10}$$

where $h_{\frac{l}{2}i}$ is the output of encoding layer representing users' diverse preferences extracted from auxiliary information. Similarly, the conditional distribution over item-specific latent features is given as,

$$p\left(V|W, D, \sigma_v^2\right) = \prod_j^M N\left(v_j | C_{out}(W, D_j), \sigma_v^2 I\right), \tag{11}$$

where D is the set of item comments. By substituting Eqs. (10) and (11) into Eq. (1), the hybrid probabilistic matrix factor model could be reformulated, and the new objective function can now be written as

$$L = \sum_{i}^{N}\sum_{j}^{M}\frac{I_{ij}}{2}(R_{ij} - U_i^T V_j)^2 + \frac{\lambda_u}{2}\sum_{i}^{N}\|U_i - h_{\frac{l}{2}i}\|_F^2 + \frac{\lambda_v}{2}\sum_{j}^{M}\|V_j - C_{out}(W, D_j))\|_F^2 + \frac{\lambda_w}{2}\sum_{k}\|W_k\|_F^2 + \frac{\lambda}{2}\sum_{l}\left(\|W_l\|_F^2 + \|V_l\|_F^2 + \|b_l\|_F^2\right), \tag{12}$$

where α is control parameter measuring the contribution of users' auxiliary information, λ_u, λ_v, λ_w and λ are weighting parameters, W_l and V_l are weights of the neurons in aSDAE component, and b_l is the bias. These parameters are estimated using maximum a posteriori (MAP) estimation [26]. Similar to [11,26], coordinate ascent algorithm is adopted to maximize variables U and V. The optimization is performed by iteratively updating one variable and fixing the rest ones. We compute the partial gradients of L with respect to u_i and v_j and set them to 0 to make the optimization. The corresponding updating equations are directly given as

$$\begin{aligned} u_i &\leftarrow (VI_iV^T + \lambda_u I_k)^{-1}(VR_i + \lambda_u h_{\frac{l}{2}i}) \\ v_j &\leftarrow (UI_iU^T + \lambda_v I_k)^{-1}(UR_j + \lambda_v C_{out}(W, D_j)), \end{aligned} \tag{13}$$

where $I_i = diag(I_{i1}, \ldots I_{iM})$ is a diagonal matrix, and I_{ij} indicates a non-empty entity in R. $R_i = (R_{i1}, \ldots, R_{iM})$ is a column vector denoting rating scores of user u_i.

After updating U and V using Eq. (13), we then update weight matrices W, W_l and V_l iteratively. W is closely related to item-specific latent features calculated using the proposed aCNN model, and W_l and V_l are closely related to user-specific latent features extracted using aSDAE component. To update these weight matrices, we first fix parameters of aCNN component and update the parameter set of aSDAE component according to the updating equations. Then, parameters of aSDAE component are fixed and we update the parameters of aCNN component iteratively. The objective function to update the rest parameters will be degenerated to below Equations if we fix the updated parameters W_l, V_l, W, U and V, written as

$$\begin{aligned} L_1 &= \frac{\lambda_v}{2}\sum_{j}^{M}\|V_j - C_{out}(W, D_j)\|^2 + \frac{\lambda_W}{2}\sum_{k}^{w_k}\|w_k\|^2 + c_1 \\ L_2 &= \frac{\lambda_u}{2}\sum_{i}^{N}\|U_i - h_{\frac{l}{2}i}\|^2 + \frac{\lambda}{2}\sum_{l}(\|W_l\|_F^2 + \|V_l\|_F^2 + \|b_l\|_F^2) + c_2, \end{aligned}$$

where c_1 and c_2 are constants. By following above equations, we can update parameters using back propagation algorithm. Therefore, U, V, W, W_l and V_l are alternatively updated until convergence or stopping criteria are satisfied. The predicted rating scores are updated by using the learnt parameter set, given as,

$$R_{ij} = u_i^T v_j = h_{\frac{l}{2}i}^T C_{out}(W, D_j) \tag{14}$$

4. Experiments

4.1. Experimental settings

For experimental evaluation, we evaluate the proposed DUPIA approach on two widely adopted real-word datasets for recommendation systems, i.e., MovieLens[1] and Amazon[2]. The original MovieLens dataset consists of 72,000 users, 10,681 movies and 10,000,054 rating scores ranging from 1 to 5. In the experiments, three sub sets are adopted which are ML-100K, ML-1M and ML-10M. The users' auxiliary information includes age, gender, occupation and zipcode. To associate user rating scores and the corresponding reviews, we also collect movies' reviews from the official movies database, IMDB[3] by following [13]. For Amazon dataset, it contains both users' ratings and products' reviews. Two sub datasets are used in the experiments which are Amazon Instant Video (AIV) and Amazon2. These two sub datasets are very sparse. The density of AIV is only 0.005%. Amazon2 contains 30,759 users, 16,515 products and 285,644 ratings with six types of products, and the average density is 0.056%. Before the experiments, several data preprocessing steps are needed which are illustrated in the next subsection.

[1] https://grouplens.org/datasets/movielens/.
[2] http://jmcauley.ucsd.edu/data/amazon/.
[3] https://www.imdb.com/.

Table 1
Statistics of the MovieLens&AIV datasets after pre-processing.

Dataset	User	Item	Rating	Density(%)
ML-100K	943	1546	94,808	6.503
ML-1M	6,040	3544	993,482	4.641
ML-10M	69,757	10,073	9,945,875	1.413
AIV	29,757	15,149	135,188	0.030

Table 2
Statistics of Amazon2 after pre-processing.

Dataset	User	Item	Rating	Density(%)
Automotive	2928	1835	20,473	0.381
office products	4905	2420	53,258	0.449
Musical instruments	1429	900	10,261	0.798
Amazon instant video	5130	1685	37,126	0.394
Patio lawn and garden	1686	962	13,272	0.818
Grocery and gourmet food	14,681	8173	151,254	0.126

4.1.1. Data pre-processing

For fairness comparison, we also preprocess the original dataset in the same way as [26] and [29]. We remove items that do not have descriptive information like comments and the statistics of the new MovieLens datasets are recorded in Table 1. For AIV dataset, we remove users who have rated less than 3 items and the statistics of the Amzon2 dataset after processing are reported in Table 2.

For users' auxiliary information, the official MovieLens dataset contains user's attributes such as age, gender, occupation and zipcode. These attributes are considered as users' auxiliary information and are encoded into a binary vector using one-hot encoding. However, there is no such data for AIV dataset. Therefore, we choose to collate top three tags that users' have labeled on movies. After preprocessing, the length of extracted features is 1822 for MovieLens-100K, 3433 for MovieLens-1M, 3969 for MovieLens-10M and 8072 on Amazon Instant Video (AIV).

For items' reviews, we take following steps. We set the maximum length of raw documents to 300 and stemming is performed on these documents. For each term, the simple TF-IDF is calculated as its feature value. Note that we also remove corpus-specific words if their document frequency is higher than 0.5. At last, we choose top 8000 distinct terms to form the vocabulary set and the rest terms are removed from the raw documents. After these preprocessing steps, the average number of terms per document is 185.67 on MovieLens-100K, 97.09 on MovieLens-1M, 92.05 on MovieLens-10M and 91.50 on Amazon Instant Video (AIV), respectively.

4.1.2. Evaluation baselines and criterion

For model performance evaluation, both baseline models as well as several state-of-the-art approaches are chosen for model comparison. Details of these approaches are illustrated as follows.

- **PMF:** Probabilistic Matrix Factorization is the baseline approach which predicts users' scores based on users' latent features and items' latent features. With these two latent features, it is assumed to be able to represent the diverse interests of users. In this approach, both users' scores as well as items' distribution are assumed to follow Gaussian distribution.
- **CTR:** Collaborative Topic Regression is a tightly coupled model which combines collaborative filtering with PMF and latent dirichlet allocation (LDA). This model tries to predict rating scores using both rating scores and document information such as reviews.
- **CDL:** Collaborative Deep Learning is a newly proposed deep Bayesian model which predicts rating scores using a carefully designed deep learning model called stacked denoising autoencoder which tries to extract users' preferences from auxiliary information.
- **ConvMF:** Convolutional Matrix Factorization is one of the state-of-the-art context-aware recommendation model which integrates convolutional neural network into PMF model. It learns textual features of reviews for the recommendations.
- **ConvMF + :** This approach extends ConvMF by using a pre-trained word embedding model, i.e., Glove [18].
- **D-Attn:** The Dual Local and Global Attention model learns latent factors of each user and each item with local and global attention based convolutional neural network. It aggregates the contribution of users' reviews and utilizes them to capture the importance of core terms contained in products' comments.

For evaluation criterion, we choose one of the most widely adopted criteria in recommendation systems which is the root mean square error (RMSE), calculated as

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{R_{ij}\in T}(R_{ij} - R'_{ij})^2}, \tag{15}$$

where R and R' represent the original and predicted rating scores, respectively.

4.1.3. Parameter setting

To train deep learning models (aSDAE and aCNN), RMSprop [15] is used in the experiments as it fits sparse data very well. To initialize model parameters, we set parameter α and learning rate of aSDAE component to 0.2 and 0.004, respectively. As the original SDAE is trained on perturbed data with a random noise, we also add noise to the users' auxiliary information. The perturbed data is generated as follows. We set the same drop out rate to 0.3 and users' auxiliary data are randomized using this drop out rate. The dimension of user-specific latent feature is set to 50. For the training of aCNN component, we set the feature dimension of input comments to 200. For the embedding layer, we use the pre-trained word embedding model to initialize the word vector. For attention and convolution layers, we empirically investigate the effect of kernel size by varying the kernel size to different values. All experiments are implemented using Tensorflow on GPU (Nvidia GeForce TitanX).

4.2. Experimental results

4.2.1. Performance comparison results

To evaluate the proposed approach, we first randomly partition the experimental datasets (MovieLens and AIV) into training, testing and validation datasets at the ratio of 8:1:1. We implement the proposed DUPIA on all sub datasets as well as the rest approaches on ML100K datasets. Due to the huge computation cost, we directly compared our approach with those results of the rest ones on the larger datasets. All experiments are repeated 5 times and the average results are recorded in Table 3.

In this table, the smaller the RMSE, the better the model performance. The last row indicates the improvement percentage that our proposed model achieved when compared with other models. It is well noticed that our proposed DUPIA achieves the best performance on all sub datasets and the model performance is increased by 1.41% on ML100K, 1.16% on ML1M, 1.35% on ML10M and 11.87% on AIV when compared with the second best model. It is also noticed that the second best model is the state-of-the-art ConvMF. It can achieve much better performance than the reset models. Moreover, the proposed DUPIA performs very well on the sparse dataset AIV, and its performance is increased by almost 12%. This indicates that the proposed approach can work very well for the "cold-start" situation. The reason might be that users' diverse preferences plus items' attractiveness can well separate different groups of users at a finer level. And these "subtle" groups of users can be used to predict rating scores more precisely.

Furthermore, we also investigate how the ratio of training data affects the model performance [23]. Therefore, we tune the density of MovieLens dataset from 20% to 80% and the corresponding RMSE values are plotted in Fig. 2. From this figure, it is well noticed that the proposed DUPIA performs constantly the best and the ConvMF is the second best one. With the increase of the ratio of training data, the RMSE of all approaches gradually decreases which is consistent with our common sense that model performance increases if more training data were used. From results of Table 3 and Fig. 2, we can conclude that the proposed DUPIA is superior to the rest approaches with respect to RMSE criterion.

To further verify the effectiveness of the proposed approach, we compare the DUPIA with the item attention based approaches on Amazon2 dataset. As is believed, the item may attract different users and these attractiveness can play an important role in predicting scores. In the literature, the state-of-art D-Attn can achieve the best model performance and therefore the experimental results are directly compared with our approach. In this experiment, both ConvMF+ and DUPIA are implemented on Amazon2 dataset and the ordinal results of D-Attn are copied to the third column of Table 4. From the results in Table 4, it is obvious that DUPIA achieves the best performance on 5 categories of products in Amazon2 dataset. For category "Musical Instruments", D-Attn is only slightly better than DUPIA, i.e., 0.8331 vs 0.8830. One possible reason is

Table 3
Evaluation results on MovieLens and AIV datasets.

Model	ML100K	ML1M	ML10M	AIV
PMF	0.9412	0.8971	0.8311	1.4118
CTR	\	0.8969	0.8275	1.5496
CDL	\	0.8879	0.8186	1.3594
ConvMF	0.9269	0.8531	0.7958	1.1337
ConvMF+	0.9280	0.8549	0.7930	1.1279
DUPIA	**0.9138**	**0.8432**	**0.7823**	**0.994**
Improve	1.41%	1.16%	1.35%	11.87%

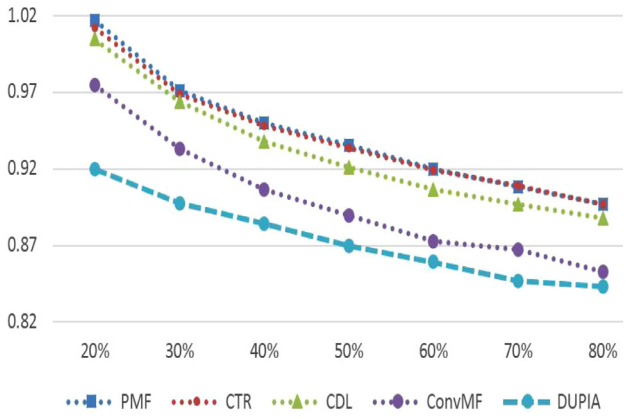

Fig. 2. Evaluation results by varying the ratio of training data.

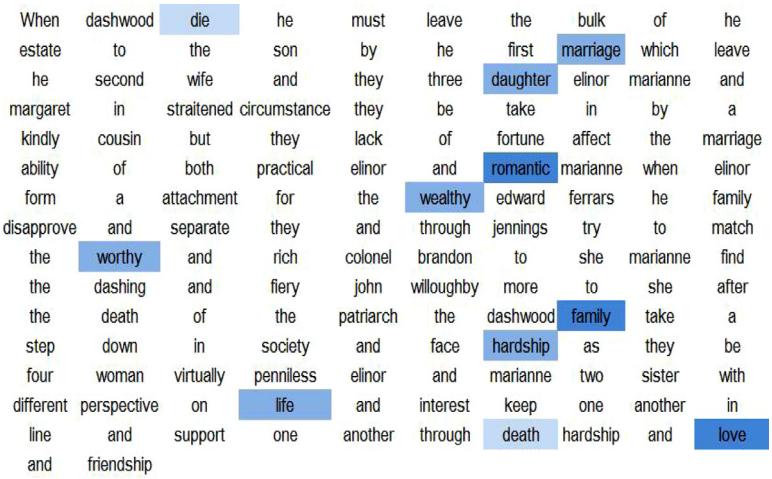

Fig. 3. Extracted feature map of movie 1 using attention based CNN.

Table 4
Evaluation results of item attention based approaches on Amazon2 dataset.

Dataset	ConvMF+	D-Attn	DUPIA
Amazon instant video	1.0575	0.9726	**0.9491**
Automotive	0.9572	0.8826	**0.8711**
Grocery and gourmet food	1.0009	0.9985	**0.9804**
Musical instruments	0.9745	**0.8331**	0.8830
Office products	0.9145	0.8479	**0.8440**
Patio lawn and garden	1.0106	0.9980	**0.9313**

that it is intuitively hard to estimate whether a user will like some specific musical instruments or not if given his age or zipcode.

4.2.2. Item attractiveness analysis

As is proposed, we believe that different persons may have different emotional tendencies when they read the same item comments. Their rating scores are partially susceptible to their emotional tendencies, i.e., item attractiveness in this paper. To further analyze how such item attractiveness can affect the rating scores, we perform the following experiments. The experiment is designed as follows. Suppose we can find two users and two movies. If user A gives 5 points on one movie but user B gives 2 points, we can conclude that user A may like this kind of movie but user B does not. For another movie, denoted as movie 2, having the same class label as movie 1, if user A gives a lower score on movie 2 but user B gives a higher score, we can conclude that these movies must have their own unique parts attracting different users.

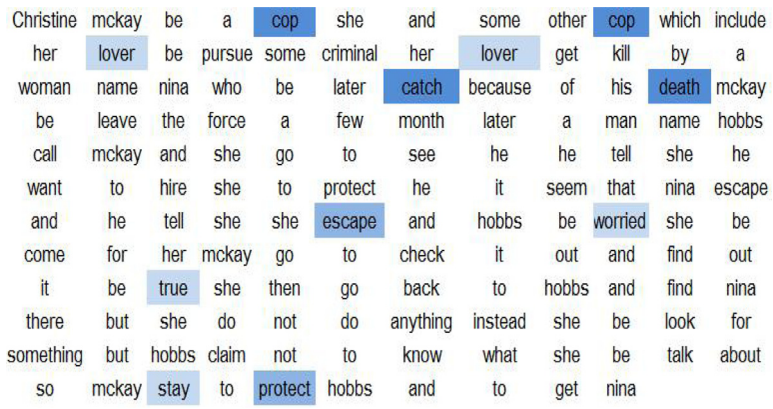

Fig. 4. Extracted feature map of movie 2 using attention based CNN.

To verify this assumption, we examine whether the core words of these two movies overlap or not for the same user. Therefore, after running the DUPIA model, we randomly choose two rows (denoting users) and the corresponding rated columns (denoting movies) from the user-item rating matrix. From these columns, we choose the columns that their labels are "Drama" for the later experiments. We extract reviews of two movies with different rating scores (by these two users). The reviews are respectively reported in Figs. 3 and 4. It is noticed that they both belong to "drama" and share some common words like "death" and "love". We then calculated the importance of each term in these reviews using Eq. (6), and we highlight the important terms in blue color. The darker blue terms are more important terms and vice versa. From Fig. 3, it is obvious that important terms are "romantic", "family", "love" and "marriage", whereas important terms in Fig. 4 are "death", "catch", "cop" and "escape". From these important terms, we know that these movies are quite different although they belong to the same class "Drama". We respectively calculate the important term vectors of movie reviews that two users give higher scores. The distance between two vectors are father than their average distance of the same class of movies that they give similar scores. This partially verifies that items' descriptive information may attract different persons.

5. Conclusions

With the prevalence of various large-scale E-commerce Websites and mobile applications, users' auxiliary information as well as items' comments can be easily collected. Naturally, deep learning based approaches are proposed to utilize either users' auxiliary information or items' textual comments to recommend items. This paper attempts to make predictions using both users' auxiliary information and items' textual comments simultaneously. In particular, a novel hybrid probabilistic matrix factorization model is proposed which tries to model users' preferences from their auxiliary information and differentiate the effect of the core terms extracted from item's comments. Particularly, two sub deep learning based components are designed for this task. And a global objective function is proposed which optimize model parameters under a unified framework. Extensive experiments are evaluated on a number of real-world datasets including 3 sub sets of MovieLens and 2 sub sets of Amazon. From the promising results, we can conclude that the proposed DUPIA is superior to both baseline models and the state-of-the-art approaches with respect to RMSE criterion. Furthermore, the proposed attention based CNN component can well capture the most attractive terms for different users.

Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

CRediT authorship contribution statement

Xiaofeng Zhang: Conceptualization, Methodology, Investigation, Writing - original draft. **Huijie Liu:** Methodology, Formal analysis, Software, Writing - original draft. **Xiaoyun Chen:** Formal analysis, Writing - original draft. **Jingbin Zhong:** Software, Validation. **Di Wang:** Validation, Writing - review & editing.

Acknowledgement

This paper is partially supported by National Science Foundation of China under Grant No. 61872108 and Shenzhen Science and Technology Program under Grant no. JCYJ20170811153507788.

References

[1] A.A. Kardan, M. Ebrahimi, A novel approach to hybrid recommendation systems based on association rules mining for content recommendation in asynchronous discussion groups, Inf. Sci. 219 (2013) 93–110.
[2] G. Attardi, DeepNL: a deep learning NLP pipeline, in: Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing, 2015, 2015, pp. 109–115.
[3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[4] J. Bobadilla, F. Ortega, A. Hernando, A. Gutirrez, Recommender systems survey, Knowl. Based Syst. 46 (1) (2013) 109–132.
[5] K. Cho, B. Van Merrienboer, D. Bahdanau, Y. Bengio, in: On the properties of neural machine translation: encoder-decoder approaches, Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (2014) 103–111.
[6] X. Dong, L. Yu, Z. Wu, Y. Sun, L. Yuan, F. Zhang, A hybrid collaborative filtering model with deep structure for recommender systems, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA., 2017, pp. 1309–1315.
[7] T. Ebesu, Y. Fang, Neural citation network for context-aware citation recommendation, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7–11, 2017, 2017, pp. 1093–1096.
[8] K. Georgiev, P. Nakov, A non-IID framework for collaborative filtering with Restricted Boltzmann Machines, in: Proceedings of the International Conference on International Conference on Machine Learning, 2013, 2013, pp. 1148–1156.
[9] T. Guo, J. Luo, K. Dong, MingYang, Differentially private graph-link analysis based social recommendation, Inf. Sci. 463–464 (2018) 214–226.
[10] M. Hong, J. J.Jung, Multi-sided recommendation based on social tensor factorization, Inf. Sci. 447 (2018) 140–156.
[11] Y. Hu, Y. Koren, C. Volinsky, Collaborative filtering for implicit feedback datasets, in: Proceedings of the Eighth IEEE International Conference on Data Mining, 2009, 2009, pp. 263–272.
[12] N. Kalchbrenner, E. Grefenstette, P. Blunsom, in: A convolutional neural network for modelling sentences, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2014) 655–665.
[13] D. Kim, C. Park, J. Oh, S. Lee, H. Yu, Convolutional matrix factorization for document context-aware recommendation, in: Proceedings of the 10th ACM Conference on Recommender Systems, in: RecSys '16, 2016, pp. 233–240.
[14] C. Li, W.K. Cheung, Y. Ye, X. Zhang, D. Chu, X. Li, The author-topic-community model for author interest profiling and community discovery, Knowl. Inf. Syst. 44 (2015) 359–383.
[15] M.C. Mukkamala, M. Hein, Variants of RMSProp and Adagrad with logarithmic regret bounds, in: Proceedings of the 34th International Conference on Machine Learning, 2017, 2017, pp. 2545–2553.
[16] M. Nickel, V. Tresp, H. Kriegel, A three-way model for collective learning on multi-relational data, in: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011, 2011, pp. 809–816.
[17] A.v.d. Oord, S. Dieleman, B. Schrauwen, Deep content-based music recommendation, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, in: NIPS'13, 2013, pp. 2643–2651.
[18] J. Pennington, R. Socher, C. Manning, GloVe: global vectors for word representation, in: Proceedings of International Conference on Empirical Methods in Natural Language Processing, 2014, 2014, pp. 1532–1543.
[19] R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization, in: International Conference on Neural Information Processing Systems, 2007, 2007, pp. 1257–1264.
[20] S. Seo, J. Huang, H. Yang, Y. Liu, Interpretable convolutional neural networks with dual local and global attention for review rating prediction, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, 2017, pp. 297–305.
[21] Y. Shi, M. Larson, A. Hanjalic, Collaborative filtering beyond the user-item matrix:a survey of the state of the art and future challenges, ACM Comput. Surv. 47 (1) (2014) 1–45.
[22] A.P. Singh, G.J. Gordon, Relational learning via collective matrix factorization, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24–27, 2008, 2008, pp. 650–658.
[23] Z. Sun, X. Zhang, Y. Ye, X. Chu, Z. Liu, A probabilistic approach towards an unbiased semi-supervised cluster tree, Knowl. Based Syst. (2019) 359–383.
[24] B. Twardowski, Modelling contextual information in session-aware recommender systems with neural networks, in: Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15–19, 2016, 2016, pp. 273–276.
[25] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[26] C. Wang, D.M. Blei, Collaborative topic modeling for recommending scientific articles, in: Proceedings of the 17th International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Diego, CA, USA, August 21–24, 2011, 2011, pp. 448–456.
[27] E. Wang, Y. Yang, K. Lou, User selection utilizing data properties in mobile crowdsensing, IEEE Trans. Mob. Comput. 490 (2019) 210–226.
[28] E. Wang, Y. Yang, J. Wu, W. Liu, X. Wang, An efficient prediction-based user recruitment for mobile crowdsensing, IEEE Trans. Mob. Comput. 17 (2018) 16–28.
[29] H. Wang, N. Wang, D. Yeung, Collaborative deep learning for recommender systems, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10–13, 2015, 2015, pp. 1235–1244.
[30] X. Wang, Y. Wang, Improving content-based and hybrid music recommendation using deep learning, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, 2014, pp. 627–636. New York, NY, USA
[31] Z. Xu, C. Chen, T. Lukasiewicz, Y. Miao, X. Meng, Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling, in: Proceedings of the ACM International on Conference on Information and Knowledge Management, 2016, 2016, pp. 1921–1924.
[32] M. Yi, Collaborative filtering, Computer Science 57 (4) (2017). 189–189
[33] P.S. Yu, P.S. Yu, P.S. Yu, Joint deep modeling of users and items using reviews for recommendation, in: Proceedings of the tenth ACM International Conference on Web Search and Data Mining, 2017, 2017, pp. 425–434.
[34] J. Zhong, X. Zhang, Wasserstein autoencoders for collaborative filtering, (2018), arXiv:1809.05662.
[35] K. Zhou, H. Zha, Learning binary codes for collaborative filtering, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, 2012, pp. 498–506.
[36] Y. Zhou, C. Huang, Q. Hu, J. Zhu, Y. Tang, Personalized learning full-path recommendation model based on LSTM neural networks, Inf. Sci. 444 (2018) 135–152.