Collaborative Multi-Level Embedding Learning from Reviews for Rating Prediction

Wei Zhang[†] **Quan Yuan**[‡] **Jiawei Han**[‡] **Jianyong Wang**[†♮]
[†]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[‡]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[♮]Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, China
zwei11@mails.tsinghua.edu.cn, {qyuan,hanj}@illinois.edu, jianyong@tsinghua.edu.cn

Abstract

We investigate the problem of personalized review-based rating prediction which aims at predicting users' ratings for items that they have not evaluated by using their historical reviews and ratings. Most of existing methods solve this problem by integrating topic model and latent factor model to learn interpretable user and items factors. However, these methods cannot utilize word local context information of reviews. Moreover, it simply restricts user and item representations equivalent to their review representations, which may bring some irrelevant information in review text and harm the accuracy of rating prediction.

In this paper, we propose a novel Collaborative Multi-Level Embedding (CMLE) model to address these limitations. The main technical contribution of CMLE is to integrate word embedding model with standard matrix factorization model through a projection level. This allows CMLE to inherit the ability of capturing word local context information from word embedding model and relax the strict equivalence requirement by projecting review embedding to user and item embeddings. A joint optimization problem is formulated and solved through an efficient stochastic gradient ascent algorithm. Empirical evaluations on real datasets show CMLE outperforms several competitive methods and can solve the two limitations well.

1 Introduction

Personalized rating prediction is a fundamental problem for recommender system and has attracted a lot of attention since Netflix Prize Challenge [Bell and Koren, 2007] was successfully held. This problem aims at leveraging historical ratings to predict users' ratings (e.g. integers from 1 up to 5) for the items they have not rated before. Among various methods for this problem, latent factor models such as matrix factorization perform better [Koren *et al.*, 2009].

As an extension of the above classical problem, personalized review-based rating prediction is newly formulated [McAuley and Leskovec, 2013] which gains extra knowledge from review text for predicting rating scores. The problem is significant as reviews can reveal the preference of users and indicate the characteristics of items, which turn out to be important references for making buying decisions. Besides, abundant of reviews can be easily acquired since they are widely available in electronic commerce companies (e.g. Amazon), video sharing websites (e.g. Youtube), and community websites (e.g. Yelp). It may enable us to achieve better prediction performance and obtain interpretable learned user and item factors.

Different from traditional sentiment analysis tasks, personalized review-based rating prediction is to predict ratings for user and item pairs when review text information only exists in historical data and is not available in prediction stage. Consequently, many sentiment analysis techniques over text cannot be exploited [Liu, 2012; Pang and Lee, 2005]. To enable effective utilization of textual content in historical data, the main challenges are how to learn useful knowledge from review text and how to adapt the learned knowledge to user and item factors to generate more accurate ratings.

Limitations of prior studies. Existing techniques [McAuley and Leskovec, 2013; Bao *et al.*, 2014; Diao *et al.*, 2014; Hu *et al.*, 2015] for this problem are often based on latent Dirichlet allocation (LDA) [Blei *et al.*, 2003] or non-negative matrix factorization (NMF) [Lee and Seung, 2000] to obtain topic factors of review text and then integrate them with matrix factorization model. However, almost all these methods suffer from at least one of the following two limitations: ignorance of word local context information and direct equivalence between user, item, and review topic factors. First, for only using document-level co-occurrence information to model words in review text, LDA and NMF will lose word order and local context information, which is a prominent issue for sentiment analysis [Wang and Manning, 2012; Johnson and Zhang, 2015]. Second, review topic factor is directly equaled to the normalized exponential transformation of user or item factor, which is not very suitable. It is intuitive because users and items cannot be fully represented by reviews due to their limited length. And not all words and topics in each review are relevant to its rating score. Direct equivalence may bring irrelevant information for rating prediction to some dimensions of embeddings.

Our solution. To address the above two limitations, we propose a novel Collaborative Multi-Level Embedding (CMLE)

model for the personalized review-based rating problem. From a whole perspective, it naturally integrates word embedding model [Mikolov *et al.*, 2013b; Le and Mikolov, 2014] with bias matrix factorization [Koren *et al.*, 2009] in a unified framework and consists of three levels (see Figure 2). In the top level, CMLE captures word local context information by word and review embedding learning, which addresses the first limitation. In the middle level, an additional projection level is proposed to project review embedding to user and item embeddings, which tightly couples them without the restricted requirement of equivalence and thus can cope with the second limitation. In the bottom level, MF is leveraged to generate ratings based on the projected user and item embeddings. In particular, we formulate a joint optimization problem to simultaneously learn word, review, user and item embeddings. An efficient stochastic gradient ascent algorithm is adopted to ensure efficiency. Without specification, embedding and factor denote the same concept in this paper.

Contributions. To sum up, the major contributions of this paper are presented as follows:

- We emphasize the two limitations existed in the previous methods for personalized review-based rating prediction problem.
- A novel model called CMLE is proposed to cope with these limitations and a joint optimization problem is formulated to learn all embeddings. To the best of our knowledge, CMLE is the first model towards integration of word embedding learning with standard latent factor model for personalized rating prediction problem.
- Experimental results on real datasets demonstrate that CMLE outperforms several previous methods for this problem and verify its benefits from overcoming the two limitations.

2 Preliminaries

2.1 Problem Formulation

Let $\mathcal{U}$ and $\mathcal{V}$ denote user and item sets, respectively. In accordance with existing studies on personalized rating prediction, we adopt $\mathbf{R}$ to represent the rating scores users assigned to items. Specifically, for $\forall u \in \mathcal{U}$ and $\forall v \in \mathcal{V}$, if u has rated v, then r_{uv} denotes the corresponding rating score. Otherwise, it is missing and needed to be predicted.

For the new problem we studied, review textual content is considered besides rating scores. Assume $\mathcal{D}$ is a collection of review text, then each review d_{uv} ($d_{uv} \in \mathcal{D}$), written by user u for item v, consists of a series of words from a vocabulary $\mathcal{W}$ and is associated with a rating score r_{uv}. Based on the above formulation, we can define the problem as follows,

Problem (Personalized Review-based Rating Prediction) *Given a user set $\mathcal{U}$, an item set $\mathcal{V}$, and a historical review text collection $\mathcal{D}$ accompanied with known entries ($r \in \tilde{\mathbf{R}}$) in a rating matrix $\mathbf{R}$, the target is to predict the scores of the missing entries ($r \in \mathbf{R} \setminus \tilde{\mathbf{R}}$) in the rating matrix.*

2.2 Motivations

As we mentioned, although integrated topic and latent factor models are capable of utilizing reviews for rating prediction, they suffer from two limitations. Inspired by the idea, we emphasize two intuitions which motivate this work.

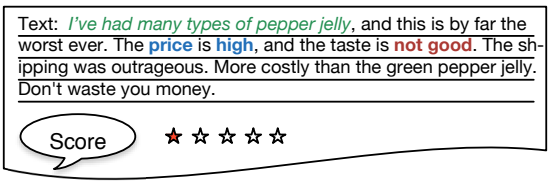


Figure 1: A simple instance (score one).

Intuition 1 *Word local context information plays an important role for rating score determination.*

Word local context information refers to the surrounding text for each word. This is a very natural intuition about how words in reviews determine rating scores. Consider the example shown in Figure 1. Generally speaking, the word "good" is a positive sentiment word, no matter which aspects of items it describes. However, the negative word "not" will change its sentiment polarity. As we know, traditional topic models only consider word co-occurrence in document level and regard all words in each document equally. Thus, local relation such as "no" for "good" cannot be captured. Another example is the word "high" shown in the review. The sentiment polarity of "high" depends on the aspect it describes. In this scenario, "high price" obviously means negative sentiment. However, if "high" is employed to modify "quality", then it conveys positive sentiment. Overall, if word local context information is not considered, then sentiment polarity may not be analyzed accurately. Inspired by the first intuition, we adopt the recently proposed embedding model (see first level of Figure 2) which naturally exploits word local context information by constructing embedding for each word based on its context words and review. Hence the proposed model can consider word local context information.

Intuition 2 *Direct equivalence between user, item, and review embeddings is a little too restricted.*

For user u and item v, we define $\mathbf{w}_u \in \mathbb{R}^{\mathcal{K}}$ and $\mathbf{w}_v \in \mathbb{R}^{\mathcal{K}}$ as their embeddings, where $\mathcal{K}$ means the dimension of embedding. d_{uv} represents the review written by user u to item v and its embedding is denoted as $\mathbf{w}_{d_{uv}}$. Direct equivalence means associating review embedding with user and item embeddings with normalized exponential transformation (e.g. $\mathbf{w}_{d_{uv}} \sim \exp(\mathbf{w}_u + \mathbf{w}_v)$). The above intuition can be simply explained from two aspects. First, some words or segments in a review may be irrelevant to the final sentiment polarity. For example, the italic segment in Figure 1 just plays a transitional role in the review and has nothing to do with the sentiment. In other words, not each topic shown in the review text contributes to the rating. Second, the average length of reviews is limited [Zhang and Wang, 2015]. Hence sparse review textual content may not fully reveal the characteristics of users and items. In summary, direct equivalence [McAuley and Leskovec, 2013; Bao *et al.*, 2014] between these embeddings still have potential to improve. Inspired by this intuition, we propose a

projection layer (see second level of Figure 2) to connect review representation to user and item representations, which relaxes the restriction of direct equivalence.

2.3 Related Models

We briefly introduce two most related models here, i.e., word embedding model and matrix factorization, which serve as two major components of CMLE.

Word Embedding Model.

Word embedding models are successful in many research fields, especially natural language processing. Two representative models, continuous bag-of-words (CBOW) and skip-gram model [Mikolov *et al.*, 2013a], are widely applied. The major difference between these two models is that CBOW leverages local context words (preceding and succeeding words) to predict its current word, while in skip-gram model, local context words are predicted by its current word.

More precisely, suppose $N_{d_{uv}}$ refers to the length of the review d_{uv}. For CBOW, its learning objective is to maximize the following log-probability,

$$\mathcal{L} = \sum_{d_{uv} \in \mathcal{D}} \sum_{k=1}^{N_{d_{uv}}} \log p(w_k | w_{k-c}^{k+c}) \tag{1}$$

where w_{k-c}^{k+c} represents a sequence of words surrounding w_k with the specified context length c which is commonly set to be 5. CBOW further defines the probability $p(w_k|w_{k-c}^{k+c})$ as,

$$p(w_k | w_{k-c}^{k+c}) = \frac{\exp(\hat{\mathbf{e}}_{w_k}^{\mathrm{T}} \sum_{-c \le j \le c, j \neq 0} \mathbf{e}_{w_j})}{\sum_w \exp(\hat{\mathbf{e}}_w^{\mathrm{T}} \sum_{-c \le j \le c, j \neq 0} \mathbf{e}_{w_j})} \tag{2}$$

where $\mathbf{e}_{w_k} (\mathbf{e}_{w_k} \in \mathbb{R}^K)$ and $\hat{\mathbf{e}}_{w_k}$ represent the input and output embedding of word w_k, respectively. It is worth noting that in this paper, we use subscripts of variables in probability expressions, like $p(w_k|w_{k-c}^{k+c})$, and specify probability computational formulas with corresponding variables, such as those shown in Equation 2.

Skip-gram model maximizes a different log-probability objective function, formulated as below,

$$\mathcal{L} = \sum_{d_{uv} \in \mathcal{D}} \sum_{k=1}^{N_{d_{uv}}} \sum_{-c \le j \le c, j \neq 0} \log p(w_j | w_k) \tag{3}$$

It is clear that the difference lies in the probability $p(w_j|w_k)$ which is defined as follows,

$$p(w_j | w_k) = \frac{\exp(\mathbf{e}_{w_k}^{\mathrm{T}} \hat{\mathbf{e}}_{w_j})}{\sum_w \exp(\mathbf{e}_{w_k}^{\mathrm{T}} \hat{\mathbf{e}}_w)} \tag{4}$$

Matrix Factorization.

Matrix factorization [Koren *et al.*, 2009] is a state-of-the-art collaborative filtering model for recommender system. Typically, rating prediction $\hat{r}_{uv}$ of user u to item v can be computed as below,

$$\hat{r}_{uv} = \mathbf{w}_u^{\mathrm{T}} \mathbf{w}_v + b_u + b_v + g \tag{5}$$

where b_u and b_v correspond to their rating biases. And g is a global rating bias.

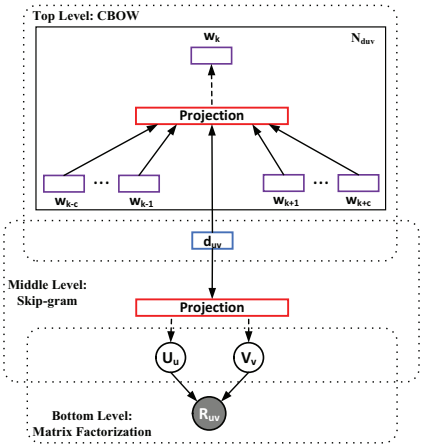


Figure 2: Graphical model of CMLE.

3 Computational Model

3.1 Model Description

To consider the two intuitions in an unified view, we propose CMLE which constructs review embedding with word embedding and further learns user and item embeddings based on review embedding and ratings. Figure 2 (rating bias terms omitted) illustrates the main idea of our model. CMLE basically consists of three levels. In the top level, it leverages the CBOW method to model each word of review text while incorporating review-level embedding. Further, review, user, and item embeddings are associated through a projection similar to skip-gram method in the middle level. Finally, the bottom level is devoted to calculating ratings through matrix factorization. In what follows, we will formalize the model level by level detailedly.

In the top level, inspired by the idea of PV-DM [Le and Mikolov, 2014], we extend the CBOW method to obtain review embedding which will later be associated with user and item embeddings. Therefore, embedding of each word in reviews depends on both its surrounding context word and review embeddings. Assume w_k belongs to review d_{uv}, then the probability of generating the word can be defined as the following softmax function,

$$p(w_k | w_{k-c}^{k+c}, d_{uv}) = \frac{\exp(\hat{\mathbf{e}}_{w_k}^{\mathrm{T}} \bar{\mathbf{v}}_{w_k})}{\sum_w \exp(\hat{\mathbf{e}}_w^{\mathrm{T}} \bar{\mathbf{v}}_{w_k})} \tag{6}$$

where $\bar{\mathbf{v}}_{w_k}$ can be calculated as below,

$$\bar{\mathbf{v}}_{w_k} = \frac{1}{2c} \sum_{-c \le j \le c, j \neq 0} \mathbf{e}_{w_{k+j}} + \mathbf{w}_{d_{uv}} \tag{7}$$

where $\mathbf{w}_{d_{uv}}$ is the embedding of review d_{uv} which summarizes the review and can be learned automatically. Because the model learning process involves simultaneously updating all the embeddings, each review embedding will be influenced by word local context information revealed by word embedding through Equation 6. As a result, review embedding can partially capture local context information, which satisfies the first intuition.

If we set α to zero then we'll have $w_u \to \odot \to r_{uv}$ w_v Just like in Neural MF

The proposed middle level with projection operation is similar to the one adopted in the skip-gram method, but their roles are fundamentally different. Through this level, CMLE can seamlessly associate review embedding with user and item embeddings. It is obvious that review embedding bridges between word embedding and user, item embedding. This is reasonable since user and item both influence construction of review text [Zhang and Wang, 2015]. To the best of our knowledge, this is the first work towards integrating word embedding learning with latent factor model for personalized rating prediction. More specifically, we define the probability of user and item embeddings based on review embedding as below,

$$\begin{aligned} p(u,v|d_{uv}) &= p(u|d_{uv})p(v|d_{uv}) \\ &= \frac{\exp(\mathbf{w}_{d_{uv}}^{\mathrm{T}}\mathbf{w}_u)}{\sum_{u'}\exp(\mathbf{w}_{d_{uv}}^{\mathrm{T}}\mathbf{w}_{u'})}\frac{\exp(\mathbf{w}_{d_{uv}}^{\mathrm{T}}\mathbf{w}_v)}{\sum_{v'}\exp(\mathbf{w}_{d_{uv}}^{\mathrm{T}}\mathbf{w}_{v'})} \end{aligned} \quad (8)$$

Compared with the formulas used in [McAuley and Leskovec, 2013; Bao *et al.*, 2014], which make user or item factor equivalent to review topic distribution through a direct normalized exponential transformation, CMLE adopts inner product of these embeddings and thus can relax the equivalence restriction.

The bottom level concentrates on the target of rating prediction. It adopts standard matrix factorization shown in Equation 5. In combination with the middle and bottom levels, we can observe that user and item embeddings capture knowledge not only from rating behaviors, but also review text information. After learning, given a new user-item pair, we can predict the rating by its user and item embeddings.

Optimization Target.

Now based on previous formulations, we derive the optimization target for later model learning. Actually, each level of CMLE has its own optimization target. For the top level, we define the target as,

$$\mathcal{L}_t = \sum_{d_{uv}\in\mathcal{D}}\sum_{k=1}^{N_{d_{uv}}}\log p(w_k|w_{k-c}^{k+c}, d_{uv}) \quad (9)$$

where $p(w_k|w_{k-c}^{k+c}, d_{uv})$ can be calculated by Equation 6. Similarly, we obtain the following log-probability for the middle level,

$$\mathcal{L}_m = \sum_{d_{uv}\in\mathcal{D}}\log p(u|d_{uv}) + \log p(v|d_{uv}) \quad (10)$$

To keep consistency with the above two targets, the objective of rating prediction in the bottom level is formulated as log-probability of Gaussian distribution as well,

$$\mathcal{L}_b = \sum_{d_{uv}\in\mathcal{D}}\log\mathcal{N}(r_{uv}-\hat{r}_{uv}, \sigma^2) \quad (11)$$

where σ^2 is set to be one for simplicity and $\hat{r}_{uv}$ can be computed through Equation 5. To avoid overfitting issue, Gaussian distributions $\mathcal{N}(\mathbf{w}_u, \lambda_U)$ and $\mathcal{N}(\mathbf{w}_v, \lambda_V)$ are usually incorporated into the above formula to play a regularization role. The precisions of Gaussian priors, λ_U and λ_V, can be regarded as hyper-parameters of regularization terms.

In summary, we can construct the final optimization target by linearly combine the above three objectives together. As learning user and item embeddings from review text is auxiliary to the main goal of predicting ratings, we utilize a weighting factor α ($\alpha \in [0,1]$) for the middle objective [Toutanova *et al.*, 2015]. It controls the relative influence of reviews and ratings on user and item embeddings. We can determine α based on the performance on validation datasets. Finally, the whole objective function can be written as,

$$\mathcal{L} = \mathcal{L}_t + \alpha\mathcal{L}_m + \mathcal{L}_b \quad (12)$$

3.2 Model Learning

We employ stochastic gradient ascent algorithm (SGA) to optimize the function shown in Equation 12, which is efficient and suitable for large-scale learning problem. For CMLE, the model parameters can be expressed as $\Theta = \{\mathbf{e}_{1:|\mathcal{W}|}, \hat{\mathbf{e}}_{1:|\mathcal{W}|}, \mathbf{w}_{1:|\mathcal{D}|}, \mathbf{w}_{1:|\mathcal{U}|}, \mathbf{w}_{1:|\mathcal{V}|}, b_{1:|\mathcal{U}|}, b_{1:|\mathcal{V}|}, g\}$. Then all model parameters can be updated based on their gradients, i.e., $\Theta^{t+1} = \Theta^t + \eta\frac{\partial\mathcal{L}}{\partial\Theta}$, where η is the learning rate and t is the current iteration number. Now the key step is to derive all gradients of the parameters. Given a training instance (u, v, d_{uv}, r_{uv}), the key gradients are calculated as below,

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\hat{\mathbf{e}}_w} &= \big(I_{[w=w_k]} - p(w|w_{k-c}^{k+c}, d_{uv})\big)\cdot\bar{\mathbf{v}}_{w_k} \\ \frac{\partial\mathcal{L}}{\partial\mathbf{e}_{w_{k+j}}} &= \frac{1}{2c}\sum_{w}\big(I_{[w=w_k]} - p(w|w_{k-c}^{k+c}, d_{uv})\big)\cdot\hat{\mathbf{e}}_w \\ \frac{\partial\mathcal{L}}{\partial\mathbf{w}_{d_{uv}}} &= 2c\sum_{k=1}^{N_{d_{uv}}}\frac{\partial\mathcal{L}}{\partial\mathbf{e}_{w_{k+j}}} + \alpha\cdot\sum_{u'}\big(I_{[u'=u]} - p(u'|d_{uv})\big)\cdot\mathbf{w}_{u'} \\ &\quad + \alpha\cdot\sum_{v'}\big(I_{[v'=v]} - p(v'|d_{uv})\big)\cdot\mathbf{w}_{v'} \\ \frac{\partial\mathcal{L}}{\partial\mathbf{w}_u} &= \alpha\cdot\big(1 - p(u|d_{uv})\big)\cdot\mathbf{w}_{d_{uv}} + (r_{uv} - \hat{r}_{uv})\cdot\mathbf{w}_v \\ \frac{\partial\mathcal{L}}{\partial\mathbf{w}_v} &= \alpha\cdot\big(1 - p(v|d_{uv})\big)\cdot\mathbf{w}_{d_{uv}} + (r_{uv} - \hat{r}_{uv})\cdot\mathbf{w}_u \end{aligned} \quad (13)$$

where $\bar{\mathbf{v}}_{w_k}$ and $\hat{r}_{uv}$ can be calculated through Equation 7 and Equation 5, respectively. For space limitation, we omit the gradients of the bias terms which are easily derived. Due to the normalization terms in Equations 6 and 8, computing these related gradients suffers a heavy cost. To speed up the learning process, we employ noise contrastive estimation (NCE) [Mnih and Teh, 2012] to approximate the softmax functions by sampling several negative words, users, and items. In this paper, we set the number of negative samples to be five and the sampling strategy is used the same as [Mikolov *et al.*, 2013b]. The above learning algorithm guarantees the cost time of each iteration grows linearly with the number of reviews and makes CMLE scalable to large datasets.

4 Experiments

4.1 Evaluation Setting

Dataset.

We conduct experiments on several real datasets which are publicly available [McAuley and Leskovec, 2013]. Based on their data sources, we call them Food, Video, and Beer. We first convert all words of the reviews into lowercase and

How about cold-start?

then remove the reviews which are too short. The basic statistics of the three datasets are: Food has 1635 users, 510 items, and 34431 reviews; Video has 2674 users, 2791 items, and 135765 reviews; Beer has 8848 users, 11190 items, and 1343985 reviews. For later comparisons, we randomly split the three datasets into train, validation, and test sets with the ratio of 7 to 1 to 2, respectively. Meanwhile, we remove users and items with less than ten occurrences in training data to ensure that users and items are associated with enough reviews. We repeat the above process five times and report the average results. We test the prediction performance in terms of two standard metrics, i.e., mean square error (MSE) and mean absolute error (MAE).

Comparison Methods.

To demonstrate the superiority of CMLE, we consider the following comparisons:

GloAve: This method simply computes the mean of historical rating scores and then uses the mean as predictions.

ItemKNN: ItemKNN [Sarwar *et al.*, 2001] is a basic item based collaborative filtering method for rating prediction.

PMF: Probabilistic matrix factorization [Salakhutdinov and Mnih, 2007] formulates matrix factorization from a probabilistic perspective without considering rating biases.

HFT: This method [McAuley and Leskovec, 2013] first proposes to utilize reviews to learn interpretable user or item representation for personalized review-based rating prediction problem, which inspires many studies later.

TopicMF: TopicMF [Bao *et al.*, 2014] is an extension of HFT which associates users and items with their corresponding reviews simultaneously under a non-negative matrix factorization framework.

JMARS: This complex model [Diao *et al.*, 2014] distinguishes aspect, sentiment, and background words in an unified framework. Its input depends on Part-of-Speech tagging results and thus we employ Stanford log-linear POS tagger [Toutanova *et al.*, 2003] here.

BMF: BMF [Koren *et al.*, 2009] is a state-of-the-art method for personalized rating prediction. Its results equal to CMLE when $\alpha = 0$ as it makes the gradients of users and items in CMLE are only related to ratings according to Equation 3.2.

To verify the rationality the first intuition, we design the following strategies which are used in the first level of CMLE:

RandomD: This strategy initializes review embedding randomly and keeps it fixed during the learning process of CMLE. Hence user and item embeddings cannot gain useful knowledge from review embedding.

RandomW: It only initializes word embedding randomly and keeps it fixed during the learning process, which means the model learns review embedding purely from rating behaviors.

LDA: This strategy initializes review embedding with its corresponding review topic distribution learned from latent Dirichlet allocation [Blei *et al.*, 2003]. As LDA ignores word local context information, it can be used as a direct comparison with our adopted strategy (**Doc2Vec**).

The following equivalence strategy is used in the middle level to indicate the rationality of the second intuition:

Equivalence: We adopt the same strategy used in [McAuley and Leskovec, 2013; Bao *et al.*, 2014] to connect review embedding to user and item embeddings. This is regarded as a comparison with the strategy (**Projection**) we proposed.

Table 1: Results of different models on rating prediction.

Method	Food		Video		Beer	
	MSE	MAE	MSE	MAE	MSE	MAE
GloAve	1.4566	0.9452	1.7431	1.0990	0.4829	0.5313
ItemKNN	0.4585	0.3907	1.3101	0.8679	0.4185	0.4725
PMF	1.0389	0.7924	1.2343	0.8492	0.3518	0.4540
HFT	0.5432	0.4511	1.2973	0.8829	0.3439	0.4403
TopicMF	0.4956	0.4283	1.2817	0.8747	0.3420	0.4386
JMARS	0.4961	0.4286	1.2430	0.8561	0.3462	0.4427
BMF	0.4271	0.3852	1.1556	0.8116	0.3391	0.4363
CMLE	0.4081	0.3738	1.1342	0.7984	0.3312	0.4311

Hyper-parameter Setting.

We tune the hyper-parameters of all methods based on their performance on validation datasets to ensure fair comparisons. For CMLE, we initialize the learning rate $\eta = 0.2$, regularization hyper-parameters to be 0.1 (same for other factor based methods such as BMF), and the relative weight $\alpha = 0.1$. All experiments are conducted with embedding dimension $\mathcal{K} = 40$. We also try other settings and find that 40 already ensures to reach stable results.

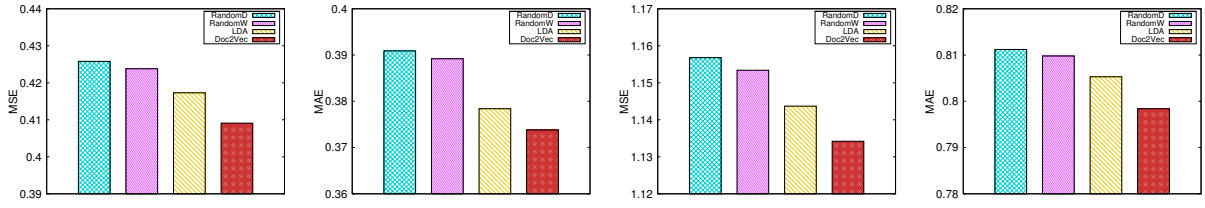
4.2 Results Analysis

Performance Comparisons.

Table 1 shows the results of CMLE and its comparisons. Overall, all the methods perform better on Food and Beer than Video. This may be explained by the fact that users' rating behaviors in video and beer domain are more diverse than that in food domain. Due to the lack of considering both personalization and review text information, GloAve performs worst among all the methods. BMF outperforms PMF consistently, which indicates the necessity of incorporating user and item rating bias for prediction. As jointly associating user and item factors with review topic, TopicMF obtains better results than HFT and its performance is roughly the same as the more complex JMARS method. Although HFT and TopicMF can provide interpretability for the learned factors, their prediction results are not as effective as BMF, which reveals their performance may be influenced by direct equivalence and is still far from satisfactory. Finally, CMLE achieves the best results among the adopted methods and the improvements are significant under t-test, which demonstrates the superiority of our model design. Besides, as the embeddings of word, user, and item lie in the same semantic space, the learned user and item embedding can be easily interpreted by the top nearest words to them (e.g calculation through inner product). Due to space limitation, the subsequent experiments do not show the results on the Beer dataset from which we can get the same conclusions as those on the other two datasets.

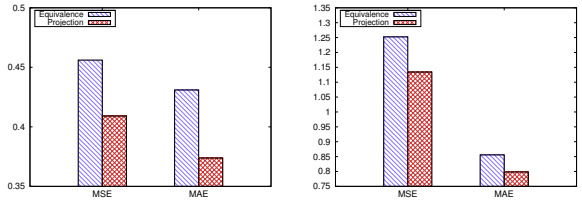
Effects of Considering Local Context Information.

Figure 3 depicts the variation of results with the different settings of embeddings. It is sensible that RandomD performs worst due to its random review embedding bringing no effective information. RandomW gains a minor improvement over RandomD because review embedding can be learned in the optimization process. Using LDA, review embedding can



(a) MSE - Food (b) MAE - Food (c) MSE - Video (d) MAE - Video

Figure 3: Comparisons of utilizing word embedding knowledge with other related methods.



(a) Integration – Food (b) Integration – Video

Figure 4: Comparisons of different integration ways.

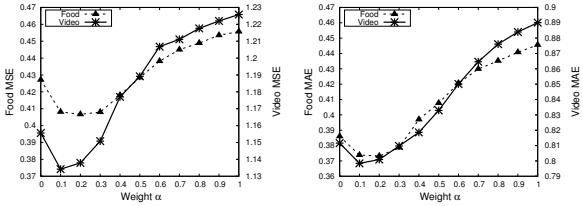


Figure 5: Influence study of relative weight α.

obtain valid knowledge from reviews and transfer them to user and item embeddings with the continual learning process. Therefore this strategy behaves significantly better than the previous two strategies. CMLE further considers word local context information to obtain more reasonable review embedding for rating prediction and thus it achieves the best results. This indicates the rationality of the first intuition.

Results of Projection versus Direct Equivalence.
We further compare the ways of associating review, user, and item embeddings. It is obvious that using the projection layer in CMLE significantly outperforms the direct equivalence strategy adopted in previous methods. This reveals direct equivalence between these embedding through a normalized exponential transformation is a little restricted. Overall, we can conclude the second intuition is reasonable.

Influence of Relative Weight α.
Finally, we study how α influences the performance of rating prediction. From Figure 5, we observe the results are optimal when α ranges from 0.1 to 0.2 and significantly better than those when $\alpha = 0$, which verifies the rationality of considering review text.

5 Related Work

Recent years have witnessed growing interest of combining personalization factor with review text information for personalized rating prediction. Personalized review-based rating prediction and personalized review-aware rating prediction are two representative problems. They differ in whether knowing review text information when prediction.

As aforementioned, many recent studies [McAuley and Leskovec, 2013; Bao *et al.*, 2014; Diao *et al.*, 2014; Zhang *et al.*, 2014; Hu *et al.*, 2015] integrate variants of topic modeling with matrix factorization for the personalized review-based rating prediction problem. [McAuley and Leskovec, 2013] first connected review topic distribution to user or item latent factor and inspired many following studies. [Bao *et al.*, 2014] further associated user and item with review text simultaneously. Furthermore, [Hu *et al.*, 2015] concluded incorporating social relations can improve prediction performance, which is analogous to our work. As we discussed, these studies suffer from the two limitations which hinders them from getting better performance. [Diao *et al.*, 2014] tried to infer aspect concentrations of users and items. Their model is a little complex and suffers from the ignorance of word local context information. [Zhang *et al.*, 2014] could also capture word local context information by first extracting aspect and sentiment words and then constructing explicit factor model. Nevertheless, their results may be influenced by the noise originated from the preprocessing step. Besides, their model is hard to incorporate user and item rating biases which are important for rating prediction.

There are fewer studies for personalized review-aware rating prediction problem. [Li *et al.*, 2014] first proposed a tensor topic model to predict the ratings for each triple of user, item, and review. [Tang *et al.*, 2015] recently formulated a deep neural network model to incorporate user and item information. As these methods require review text information known when prediction, they cannot be easily adapted to the problem we studied.

6 Conclusion and Future Work

In this paper, we have studied the problem of personalized review-based rating prediction. We emphasize the two limitations of related methods, i.e., ignorance of word local context information and direct equivalence between user, item, and review embeddings. To cope with these limitations, we propose a novel model called CMLE. To the best of our knowledge, it is the first study to integrate the word embedding model with matrix factorization for personalized rating prediction. The experimental results on real datasets demonstrate CMLE is effective and can solve the two limitations well.

For future work, several interesting directions can be ex-

plored. First, utilizing phrase embedding [Yu and Dredze, 2015] based on mined phrases [Liu *et al.*, 2015] instead of word embedding may be beneficial. Second, recent deep learning approaches [Socher *et al.*, 2013] with more ability of modeling text may be applied to the studied problem.

Acknowledgments

We thank the anonymous reviewers for their valuable and constructive comments. This work was supported in part by National Basic Research Program of China (973 Program) under Grant No. 2014CB340505, National Natural Science Foundation of China under Grant No. 61532010 and 61272088, Tsinghua University Initiative Scientific Research Program under Grant No.20131089256, the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). Besides, Wei Zhang is supported by China Scholarship Council.

References

[Bao *et al.*, 2014] Yang Bao, Hui Fang, and Jie Zhang. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *AAAI'14*, pages 2–8, 2014.

[Bell and Koren, 2007] Robert M. Bell and Yehuda Koren. Lessons from the netflix prize challenge. *SIGKDD Explorations*, 9(2):75–79, 2007.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[Diao *et al.*, 2014] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *KDD'14*, pages 193–202, 2014.

[Hu *et al.*, 2015] Guangneng Hu, Xinuu Dai, Yunya ong, Shujian Huang, and Jiajun Chen. A synthetic approach for recommendation: Combining ratings, social relations, and reviews. In *IJCAI'15*, pages 1756–1762, 2015.

[Johnson and Zhang, 2015] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. In *NAACL'15*, pages 103–112, Denver, Colorado, May–June 2015.

[Koren *et al.*, 2009] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

[Le and Mikolov, 2014] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML'14*, pages 1188–1196, 2014.

[Lee and Seung, 2000] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS'00*, pages 556–562, 2000.

[Li *et al.*, 2014] Fangtao Li, Sheng Wang, Shenghua Liu, and Ming Zhang. SUIT: A supervised user-item based topic model for sentiment analysis. In *AAAI'14*, pages 1636–1642, 2014.

[Liu *et al.*, 2015] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. Mining quality phrases from massive text corpora. In *SIGMOD'15*, pages 1729–1744, 2015.

[Liu, 2012] Bing Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[McAuley and Leskovec, 2013] Julian J. McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys'13*, pages 165–172, 2013.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS'13*, pages 3111–3119, 2013.

[Mnih and Teh, 2012] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. In *ICML'12*, pages 1751–1758, 2012.

[Pang and Lee, 2005] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL'05*, pages 115–124, 2005.

[Salakhutdinov and Mnih, 2007] Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *NIPS'07*, pages 1257–1264, 2007.

[Sarwar *et al.*, 2001] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW01*, pages 285–295, 2001.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP'13*, pages 1631–1642, 2013.

[Tang *et al.*, 2015] Duyu Tang, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *ACL'15*, pages 1014–1023, 2015.

[Toutanova *et al.*, 2003] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL'03*, 2003.

[Toutanova *et al.*, 2015] Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *EMNLP'15*, pages 1499–1509, 2015.

[Wang and Manning, 2012] Sida I. Wang and Christopher D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *ACL'12*, pages 90–94, 2012.

[Yu and Dredze, 2015] Mo Yu and Mark Dredze. Learning composition models for phrase embeddings. *TACL*, 3:227–242, 2015.

[Zhang and Wang, 2015] Wei Zhang and Jianyong Wang. Prior-based dual additive latent dirichlet allocation for user-item connected documents. In *IJCAI'15*, pages 1405–1411, 2015.

[Zhang *et al.*, 2014] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR'14*, pages 83–92, 2014.