Aspect-Similarity-Aware Historical Influence Modeling for Rating Prediction

Ryo Shimura Shotaro Misawa Masahiro Sato
Tomoki Taniguchi Tomoko Ohkuma
Fuji Xerox Co., Ltd.
{shimura.ryo,misawa.shotaro,masahiro.sato,
taniguchi.tomoki,ohkuma.tomoko}@fujixerox.co.jp

Abstract

Many e-commerce services provide customer review systems. Previous laboratory studies have indicated that the ratings recorded by these systems differ from the actual evaluations of the users, owing to the influence of historical ratings in the system. Some studies have proposed using real-world datasets to model rating prediction. Herein, we propose an aspect-similarity-aware historical influence model for rating prediction using natural language processing techniques. In general, each user provides a rating considering different aspects. Thus, it can be assumed that historical ratings provided considering similar aspects to those of later ones will influence evaluations of users more. By focusing on the review-topic similarities, we show that our method predicts ratings more accurately than the previous historical-inference-aware model. In addition, we examine whether our model can predict "intrinsic rating," which is given if users were not influenced by historical ratings. We performed an intrinsic rating prediction task, and showed that our model achieved improved performance. Our method can be useful to debias user ratings collected by customer review systems. The debiased ratings help users to make decision properly and systems to provide helpful recommendations. This might improve the user experience of e-commerce services.

1 Introduction

Currently, many e-commerce services like Amazon provide customer review systems (CRS). CRSs retrieve user feedbacks, which contain mainly ratings and reviews, shared across entire users, enabling (1) subsequent users to help decide whether to purchase the items and (2) the systems to make recommendations for items.

Previous studies have shown that historical ratings presented by a CRS can create historical influence (Adomavicius et al., 2016). In this study, "historical influence" refers to a phenomena that historical ratings make users give ratings apart from their natural evaluation. According to the previous work, users tend to give higher ratings to items presented with a high average historical rating. Such influence affects the unbiased purchase decisions of subsequent users, and provides inaccurate and unhelpful recommendations. Therefore, it is important that the recommender system estimates the "intrinsic rating", which is given if users were not influenced. Recently, some studies have proposed historical-influence-aware rating prediction models (Wang et al., 2014; Liu et al., 2016; Zhang et al., 2019). In particular, Zhang et al. (2019) found that the subsequent rating of an item correlates with the average historical ratings at the time of evaluation. They concluded that such correlation patterns could be described by an assimilation-contrast theory (Anderson, 1973). Subsequent users tend to provide ratings according to the average historical rating when it is close to their intrinsic ratings (assimilation); conversely, they provide ratings against the average historical rating when it differs from their intrinsic ratings (contrast). The proposed model is called the historical-influence-aware latent factor (HIALF) model, and it has achieved significant improvements in rating prediction. In addition, the model can be used to estimate intrinsic ratings.

Proceedings of the Workshop on Natural Language Processing in E-Commerce (EComNLP), pages 76–86
Barcelona, Spain (Online), Dec 12, 2020.

We assume that evaluators are influenced significantly by the historical ratings given under aspects similar to theirs. In general, different users focus on different aspects of an item and the rating is given based on these aspects. For example, a user concerned more about the color of an item may provide a different rating for the item than a user concerned more about the price. Therefore, it is natural that the strength of the influence from each historical rating depends on the aspects under which the rating was provided. Consequently, evaluators are likely to provide ratings higher or lower than intrinsic ratings, when the historical rating is high or low, respectively.

In this paper, we conduct preliminary analyses to confirm that our assumption is appropriate. Then, we propose an aspect-aware historical influence model for rating prediction using natural language processing techniques (Figure 1 (A)). We apply topic modeling to extract aspects from reviews, and calculate the similarities in the aspects. The calculated aspect similarity is used to weight the corresponding historical ratings. The weighted ratings are then aggregated to integrate the matrix factorization (MF) model (Koren et al., 2009). We conduct experiments to show that our model outperforms HIALF and MF in rating prediction on four real-world datasets. Additionally, to examine whether our model can predict intrinsic ratings, we evaluate our model on the additional task of predicting the first rating for each item. Our results demonstrate that the proposed model predict the intrinsic ratings of users accurately.

Contributions: The contributions of this work are summarized as follows.

- We develop an aspect-similarity-aware historical influence model using reviews.
- We show that our model can estimate the intrinsic ratings of users more accurately.
- Our results indicate that users are strongly influenced by the historical ratings provided under aspects similar to theirs.

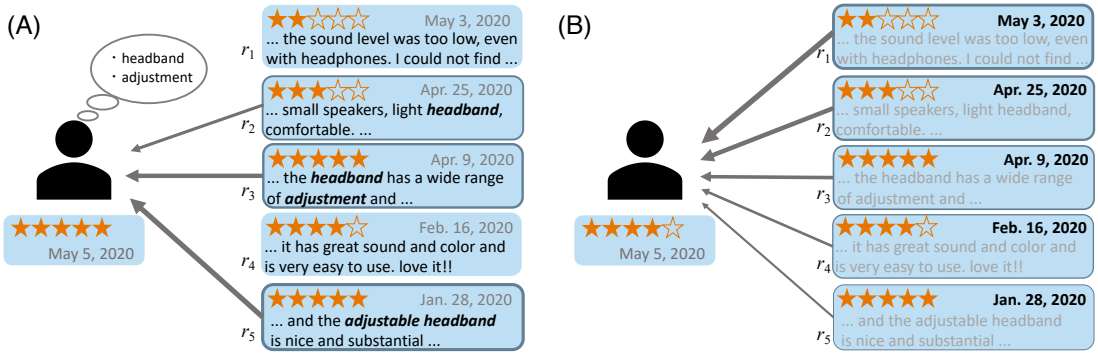

Figure 1: Illustration of historical influence models. A user purchased headphones and rated it. (A) Model where users are influenced by ratings whose review was written under aspects similar to those of the evaluator (B) Model where users are influenced more by recent ratings used in HIALF. For example, in illustration (A), evaluator u focuses on the "headband" and "adjustment" of the item. Thus, the evaluator will be influenced significantly by ratings r_2, r_3 and r_5, because their reviews mention "headband" and "adjustment" of the item.

2 Related Works

Laboratory experiment on historical influence: Adomavicius et al. (2016) showed that users may fail to evaluate an item properly when they preliminarily observe system-predicted or averaged ratings. For example, even when comparing items of the same quality, users tend to give higher ratings to the one presented with a higher average historical rating.

Modeling historical influence: Some studies have modeled historical influence in real-world datasets (Wang et al., 2014; Liu et al., 2016; Zhang et al., 2019). Wang et al. (2014) developed Herding Effect Aware Rating Dynamics Model (HEARD) introducing generative model. Although the model calculates the distribution of next rating from historical ratings to model dynamics of rating growth, it does not

take user preference into account. By contrast, Liu et al. (2016) and Zhang et al. (2019) attempted to model influence from historical ratings on subsequent users combining with MF. Liu et al. preliminarily analyzed the real-world datasets and found that the next rating correlates with the average of the historical ratings of the item. Then, they proposed the model which has the term considering the average and number of historical ratings. Similarly, Zhang et al. analyzed the other datasets and concluded that such correlation patterns can be explained by "Assimilate-Contrast" theory in psychology, based on which they developed HIALF and showed that the model outperforms previous methods. Although these methods have considered historical influence, they ignore reviews. Our model is on top of these works, and our proposed method uses reviews to model aspect-similarity-aware historical influence.

Rating prediction with reviews: Many works use reviews to improve accuracy, in particular, some of them utilize the MF model (McAuley and Leskovec, 2013; Ling et al., 2014; Tan et al., 2016; Zheng et al., 2017; Chen et al., 2018; Li et al., 2019b). They differ from our work in using reviews to model user-preferences and item-features. Our method utilizes reviews not for user/item modeling directly, but for modeling historical influence.

Modeling social influence: Recent works have studied social influence, which is another type of influence on evaluation (Ye et al., 2012; Guo et al., 2014; Li et al., 2019a; Li et al., 2019b; Wu et al., 2019a; Wu et al., 2019b). These works model social influence under explicit user-user networks such as friends, and trust. However, some of CRSs like Amazon does not have explicit network. Mukherjee and Guennemann (2019) proposed the model GhostLink, which can infer implicit user-user network from only timestamped reviews of the users. GhostLink use echoed/copied topics of reviews as an indication of influence. They showed that they can predict ratings accurately using the inferred influence network. There is a difference that while we focus on the historical influence from ratings and reviews, GhostLink is motivated to infer the user-user networks to find out who-influences-whom relationships.

3 Preliminary Analysis

Zhang et al. (2019) investigated the relationship between subsequent ratings and historical ratings at the point of each evaluation. If users provide completely uninfluenced feedbacks, there should be no correlation between subsequent ratings and historical ratings. They obtained relationships between the subsequent rating and historical ratings, and evaluated the relationships using the slopes of fitted lines in the points obtained. This is discussed in detail in Section 3.2.

Following Zhang et al., we performed preliminary analyses to confirm that users are influenced by the ratings of user feedback considering similar aspects. Specifically, we use restricted subsets of user feedbacks for averaging, instead of subsets of all historical ratings as in the study of Zhang et al. Each subset is formulated to have similar aspects for each target user feedback. Then, we plot the target ratings against the average of the subsets. Lastly, we compare the slopes of the fitted lines, and examine whether these slopes change depending on the subset used for averaging.

3.1 Datasets

We used four datasets of different categories from the Amazon dataset (He and McAuley, 2016). These datasets include reviews and ratings from May 1996 to July 2014. From these datasets, we extracted user IDs, item IDs, 1-5-star(s) ratings, free-text reviews, and timestamps. The statistics of the datasets are summarized in Table 1.

	# items	# users	# ratings&reviews
Movies and TVs	208,321	2,088,620	4,607,047
Electronics	498,196	4,261,096	7,824,482
Clothing, Shoes and Jewelry	1,503,384	3,117,268	5,748,920
Books	2,370,585	8,026,324	2,507,155

Table 1: Summary of the Amazon datasets.

3.2 Measurement Procedure

First, we describe the details of the analyses conducted by Zhang et al. Items with overall average ratings in the range of 2.9 to 3.1 were used. A prior expectation $e_{i,n}$ is calculated as

$$e_{i,n} = \frac{1}{|\mathcal{H}_{i,n}|} \sum_{r \in \mathcal{H}_{i,n}} r, \tag{1}$$

where $\mathcal{H}_{i,n}$ denotes the historical ratings of $r_{i,n}$: $\{r_{i,1}, r_{i,2}, \ldots, r_{i,n-2}, r_{i,n-1}\}$. The prior expectations are rounded off to one decimal place. For the set of the pairs of $(r_{i,n}, e_{i,n})$, a binning operation is performed by each $e_{i,n}$. As a result, we obtain bins of $\{1.0, 1.1, \ldots, 4.9, 5.0\}$, and each bin contains a set of the next ratings $\{\dot{r}_1, \ldots, \dot{r}_{N_e}\}$ given by different users under a prior expectation e. These next ratings are averaged within each bin of the prior expectation e as

$$\bar{r}_e = \frac{1}{N_e} \sum_{k=1}^{N_e} \dot{r}_k, \tag{2}$$

where N_e denotes the number of next ratings contained in the bin of the prior expectation e. Finally, the prior expectations e and average next ratings $\bar{r}_e$ are plotted, and the Pearson correlation coefficient and slopes of the fitted lines are calculated. A linear regression model is used for the fitting.

In the original procedure, all historical ratings of $r_{i,n}$ are used (Eq. 1). Here, we consider using subsets for calculating the prior expectations, instead of using $\mathcal{H}_{i,n}$. We adopt the following factors to extract subsets for each user feedback:

- *Random*: Randomly selecting ten user feedbacks.
- *Sentence-Similarity*: Selecting the ten most similar user feedbacks. We use the bag-of-words of TF-IDF to measure sentence similarity for simplicity.

An issue of concern is that some undesirable positive correlations may occur in cases where the target reviews contain words, such as "good," "great," "bad," and "terrible," that directly express the quality of the item. In these cases, the user feedback extracted based on similarity also contains such words, and has positive correlations. However, these are not derived from historical influence.

Consider this, we also apply the same procedure for the future ratings of $r_{i,n}$. The analyses using future ratings are expected to capture only positive correlations that are not related to historical influences. Therefore, by *subtracting* the effects of the future ratings from those of the historical ratings, we can measure the historical influence from each subset without undesirable positive correlations. In future ratings analyses, we extract ratings from $\{r_{i,n+1}, r_{i,n+2}, \ldots, r_{i,N_i}\}$ for a target rating $r_{i,n}$, where N_i denotes the number of ratings of an item i in the datasets.

3.3 Results

In Table 2, we compared the Pearson correlation coefficients and slopes for each factor. To avoid noise owing to the low sample size, we fitted lines to the points with prior experiments in range of 2.0 to 4.0. There were strong positive coefficients for the *Sentence-Similarity* in four datasets in both the historical and future ratings analyses. This indicates that linear regression is well-fitted. A comparison of the slopes shows that *Random* exhibits mostly flat slopes for both the historical and future rating analyses on the four datasets. By contrast, *Sentence-Similarity* exhibits positive slopes.

Sentence-Similarity exhibited high slopes on the four datasets in the historical-ratings analyses; conversely, the slopes decreased in the future ratings analyses. This implies that there may be cases where the target reviews contain words expressing quality, as discussed previously. However, the slopes of the historical-ratings analyses are higher than those of the future-ratings analyses. Therefore, we concluded that *Sentence-Similarity*, which includes aspects similarity, might actually relate to the historical influence.

	Slope		Pearson Correction Coefficient	
	Random	*Sentence-Similarity*	*Random*	*Sentence-Similarity*
Movies	0.037 (0.013)	0.718 (0.428)	0.502 (0.176)	0.997 (0.995)
Electronics	-0.028 (-0.042)	0.763 (0.665)	0.422 (0.771)	0.997 (0.996)
Clothing	-0.093 (-0.116)	0.629 (0.531)	0.752 (0.874)	0.987 (0.983)
Books	-0.037 (-0.018)	0.549 (0.407)	0.398 (0.197)	0.995 (0.989

Table 2: Results of the historical-ratings (Upper lines) and future-ratings (Lower lines, within parentheses) analyses.

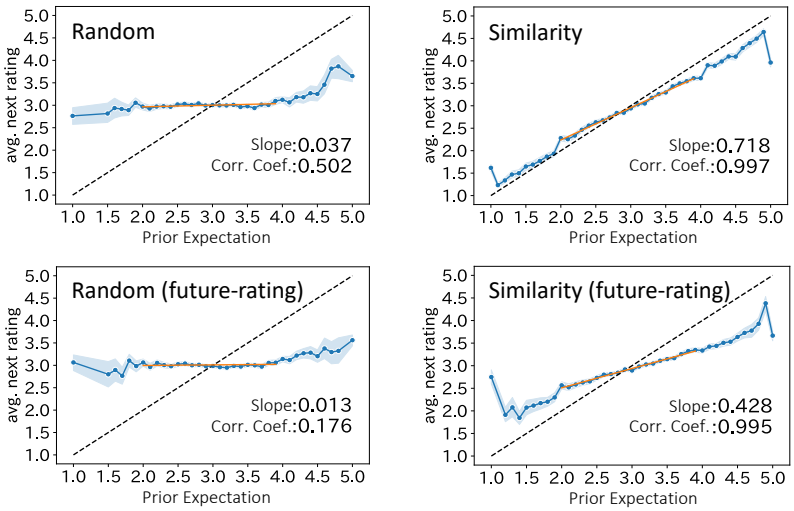

Figure 2: Plot of the prior expectations and average of next ratings in the Movies&TVs dataset. The x-axis represents the bins of prior expectation, and the y-axis represents the average of the next ratings.

4 Base Models

4.1 Biased Matrix Factorization Model

The biased MF model proposed by Koren et al. (2009) calculates an interaction between a user preference and an item feature as

$$\hat{r}_{u,i}^{\mathrm{MF}} = \mu + b_u + b_i + p_u^\top q_i. \tag{3}$$

Here, μ, b_u, and b_i denote the overall average, user bias, and item bias, respectively. p_u and q_i represent the l-dimensional user preference vector and item feature vector, respectively. This model has been widely used in rating prediction.

4.2 Historical Influence Aware Latent Factor Model

HIALF (Zhang et al., 2019) is designed under the assumption that the quality perceived by the users and prior expectations primarily form their ratings. The experienced quality represents the intrinsic evaluation of the users for item[1]. Prior expectation is defined as the average of the historical ratings that users observe prior to evaluation. Zhang et al. showed that user ratings tend to differ from the experienced quality because of being influenced. This indicates that the difference between the quality experienced by users and prior expectation is a signal of being influenced. Thus, user ratings will be

[1]Note that the "experienced quality" and "intrinsic rating" are different in this paper. $s_{u,i}$ in Eq. 4 denotes the former, which is defined in Zhang et al. (2019). The later is denoted by $b_u + s_{u,i}$ in this paper.

expressed as the sum of the experienced quality and influence based on this difference. HIALF predicts the n-th rating of i by u as:

$$r_{u,i,n}^{\mathrm{HIALF}} = \underbrace{b_u + s_{u,i}}_{\mathrm{MF}} + \underbrace{\alpha_u f\left(|\mathcal{H}_{i,n}|\right) \boldsymbol{\beta}\left(e_{i,n} - s_{u,i}\right)}_{\text{Historical influence}}. \tag{4}$$

Here, $s_{u,i}$ and $e_{i,n}$ denote the experienced quality and prior expectation, α_u denotes the likelihood of a user u being influenced, and $\mathcal{H}_{i,n}$ denotes the set of historical ratings before u provided the rating $r_{i,n}$. The experienced quality is calculated as $s_{u,i} = \hat{r}_{u,i}^{\mathrm{MF}} - b_u$. The prior expectation $e_{i,n}$ is calculated by the recency-weighted average of historical ratings as:

$$e_{i,n} = \frac{\sum_{k=1}^{n-1} \xi(n-k) \cdot r_{i,k}}{\sum_{k=1}^{n-1} \xi(n-k)}, \text{ where } \xi(d) = \exp(-\gamma * d). \tag{5}$$

The weights of each historical rating increase depending on the recency. This is based on the idea that users will focus more on recent historical ratings. In fact, Zhang et al. showed that the recency-weighted average is better than the uniform average.

Function $f(m)$ is sigmoid-form function modeling the magnifying impact of historical ratings with the size m as:

$$f(m) = \frac{a}{1 + \exp\left(-bm\right)} - \frac{a}{2}, \tag{6}$$

where m denotes the size of the historical ratings, and a, b are the learning parameters. The overall effect of the historical influence will be strong when m is large.

Function $\boldsymbol{\beta}(x)$ is a bias curve representing the assimilate-contrast effect of $x = e_{i,n} - s_{u,i}$. Non-parametric kernel regression is used to fit a set of samples $\{(g_l, v_l)\}_{l=1}^{L}$ as:

$$\boldsymbol{\beta}(x) = \frac{\sum_{l=1}^{L} w\left(x, g_l\right) \cdot v_l}{\sum_{l=1}^{L} w\left(x, g_l\right)}, \text{ where } w(x, g_l) = \exp\left(-\kappa \left(x - g_l\right)^2\right). \tag{7}$$

Here, $\{g_1, g_2, \ldots, g_{L-1}, g_L\}$ are fixed to $\{-4, -3.5, \ldots, 3.5, 4.0\}$ in order, and $\{v_1, \ldots, v_L\}$ are the learning parameters. κ is a hyperparameter controlling the smoothness of the function. The learned curve is expected to be formed as $\boldsymbol{\beta}(x)$ grows where $|x|$ is small (assimilation); otherwise, it declines (contrast).

The original MF model (Eq. 3) cannot discriminate whether a high rating is due to the intrinsic rating or historical influence. By introducing the historical influence term, the MF term in HIALF can learn intrinsic features apart from historical influence.

5 Proposed Model

We propose the model under the assumption that the strength of the historical influence depends on aspects similarity between an evaluator and users who gave historical ratings. We use a topic model to vectorize the review, and the extracted topic is regarded as an aspect of the review. Then, we calculate the similarity between the review topics of the evaluator and those of the historical users. We use this similarity as the aspect similarity, because user reviews are considered to reflect their aspects. By considering aspect similarity, the proposed method is expected to model the historical influence accurately.

Our model is an extension of HIALF. The major difference lies in the calculation of prior expectation. In HIALF, the historical ratings and recency are used in the calculation. By contrast, the proposed model uses historical ratings and review topics. Concretely, we calculate the aspect similarity between the evaluator and each historical user, and then, the prior expectation only among the ratings with high aspect similarities. The model is described as follows:

$$\hat{r}_{u,i} = b_u + s_{u,i} + \alpha_u f\left(|\mathcal{H}_{i,n}|\right) \boldsymbol{\beta}(e_{u,i}^{\mathrm{simi}} - s_{u,i}). \tag{8}$$

Here, α_u represents the likelihood of the user u being influenced, and function $f(\cdot)$ and $\boldsymbol{\beta}(\cdot)$ models the magnifying effect and the bias curve following HIALF, respectively (Eq. 6, 7). $e_{u,i}^{\mathrm{simi}}$ denotes the

aspect-aware prior expectation using the review-topic similarity. We detail the method in the following subsection.

5.1 Aspect-aware Prior Expectation

We introduce similarity-weighted aggregation method. To calculate the aspect-aware prior expectation $e^{\text{simi}}_{u,i}$, we extract top-k similar reviews and aggregate them as:

$$e^{\text{simi}}_{u,i} = \frac{\sum_{(d',r')\in\mathcal{D}^{\text{simi}}_k} r' \text{simi}^w(d', d_{u,i})}{\sum_{(d',r')\in\mathcal{D}^{\text{simi}}_k} \text{simi}^w(d', d_{u,i})}, \tag{9}$$

where $\mathcal{D}^{\text{simi}}_k$ denotes the subset of pairs of top-k similar review-topics vectors and corresponding ratings, w denotes scaling factor of similarity weighting, and $d_{u,i}$ denote the review-topics vector for item i by user u, respectively. In our paper, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is used for topic modeling and cosine similarity is used for similarity calculation. The prior expectation is designed to give large weight to historical ratings which seem to be given under similar aspects.

5.2 Objective Function

To learn the model, we define objective function as:

$$\mathcal{L}(\Theta) = \sum_{(u,i)\in\mathcal{T}} (r_{u,i} - \hat{r}_{u,i})^2 + \lambda_l(\|p_u\|^2 + \|q_i\|^2 + b_u^2 + b_i^2) + \lambda_\alpha \alpha_u^2 + \lambda_f(a^2 + b^2) + \lambda_\beta \left(\sum_{l=1}^{L} v_l^2\right). \tag{10}$$

Here, Θ denotes the learning parameters $p_u, q_i, b_u, b_i, \alpha_u, a, b$, and $\lambda_l, \lambda_\alpha, \lambda_f, \lambda_\beta$ are regularization hyperparameters for Θ. $\mathcal{T}$ denotes the set of user-item pairs in training data, and $r_{u,i}$ is the ground truth of the rating of item i by user u.

6 Experiments

We conducted two tasks: ordinay rating prediction (*Task 1*) and intrinsic rating prediction (*Task 2*). *Task 1* is widely used to evaluate recommender systems. In *Task 1*, we investigated the effect of top-k and w in Eq. 9. *Task 2* is where a model predicts the first ratings of every item. The first ratings can be used as the ground truth for the intrinsic ratings of users, because users who provided the first rating were not exposed to any historical influence. In *Task 2*, the ratings are predicted only from pre-trained users or item-features. The models are evaluated based on the ability to distinguish intrinsic ratings from historical influence. If a model successfully learns the intrinsic user-features, it will be able to accurately estimate the intrinsic ratings of users.

6.1 Experimental Settings

Dataset preprocessing: We preprocessed the four datasets described in Section 3. We preprocessed the datasets following (Zhang et al., 2019). We started by removing items with less than 75 ratings. Then, we extracted users with less than 50 ratings, and merged them into one *pseudo-user*. The pseudo-user is treated in the same manner as other users. This process aims to remove users whose data are too small to learn latent factors without reducing the historical ratings. The statistics of the preprocessed datasets are summarized in Table 3.

	# items	# users	# ratings&reviews
Movies and TV	11,194	2,311	3,079,522
Electronics	17,727	400	4,860,410
Clothing, Shoes and Jewelry	9,623	2	1,677,798
Books	42,462	9,401	9,321,929

Table 3: Summary of the statistics of preprocessed datasets.

Baseline models: We compared our model with the following baseline models in *Task 1*:

(A) *Biased-MF* (Koren et al., 2009): Classical latent factor model (Eq. 3).

(B) *HIALF* (Zhang et al., 2019): The model considering the historical influence but not utilizing reviews. Our model is based on this model.

(C) *Simi-avg.*: As baseline method, we predict ratings without LF integration, that is, $\hat{r}_{u,i} = e^{\mathrm{simi}}_{u,i}$.

In *Task 2*, we compared our proposed model with *HIALF*.

Evaluation metric: The root mean square error (RMSE) was used for performance evaluation. It was calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{S}|} \sum_{(u,i)\in\mathcal{S}} (r_{u,i} - \hat{r}_{u,i})^2}, \tag{11}$$

where $\mathcal{S}$ denotes the set of user-item pairs in the test set. $r_{u,i}$ and $\hat{r}_{u,i}$ denote the ground-truth and predicted ratings of item i by user u, respectively. Note that in the test set in *Task 2*, the historical influence term in Eq. 8 (and Eq. 4) is equal to 0. This is because the test data have no historical ratings or reviews. This is equivalent to the following:

$$\hat{r}_{u,i} = \mu + b_u + b_i + p_u^\top q_i, \tag{12}$$

where b_u, b_i, p_u, and q_i are the learned parameters of *HIALF* and the proposed model.

Model training: In *Task 1*, we split the dataset into test (the last 25 ratings), validation (the last 50-26 ratings), and training (the rest of ratings). We used a validation set to tune the hyperparameters, and performance evaluation was conducted on the test set. For all models, we used $l = 5$ for the number of dimensions of the user preference and item feature vectors. The learning rate was searched in [0.005, 0.01, 0.05 0.1] for each model. For *Biased-MF*, the regularization parameter was searched in [0.005, 0.01, 0.05, 0.1]. For *HIALF*, we searched for the best hyperparameters in the range described by Zhang et al. (2019). For our model, we first searched for w and k in [0.0, 1.0, 3.0, 5.0, 7.0, 9.0] and [5, 10, 30, 50, 100], respectively. Then, four regularization parameters were searched in [0.000001, 0.00001 0.0001, 0.001, 0.01, 0.1]. We tokenized the review-sentences, and removed stop-words using NLTK (Bird and Klein, 2009). Then, reviews were vectorized by LDA using gensim (Řehůřek and Sojka, 2010). We used 10 for the number of topics according to the results of McAuley and Leskovec (2013). The stochastic gradient descent algorithm was used for optimization.

In *Task 2*, the datasets were split into test (first rating of every item) and training (the rest of ratings). We trained the model with the hyperparameters tuned in *Task 1*.

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

6.2 Results and Discussions

Effect of w and k: We focus on the hyperparameters w and k in Eq. 9. Figure 3 shows RMSE in the Books dataset at various w and k. Here, *Simi-avg.* was used. We can see that $w = 5.0$, and $k = 30$ are the best values. The RMSE tends to improve at $w > 0$ when $k \geq 30$, indicating that weighing with the similarity is effective. Weighing is not effective when $k = 5$ or 10, because all extracted reviews have high similarities at these values. Additionally, the RMSE degrades when $k \geq 50$. This is because if we extract historical reviews that are too large, there may be low-similarity reviews.

Results of rating prediction in Task1: Table 4 illustrates the results of the rating prediction. Our model outperforms almost all baseline models. Additionally, we confirm that integrating LF (Table 4 (D)) improves the RMSE over *Simi-avg.* (Table 4 (C)). In the Clothing dataset, *Simi-avg.* performs better than *Ours*. This is because the LF model is not effective for pseudo-users, and there is only one non-pseudo-user in the Clothing dataset. These results indicate that we succeeded in integrating LF and the historical influence model.

Note that *Ours* and *Simi-avg.* use review textual information of target users. Since user ratings and reviews are usually provided at the same time, recommender systems cannot use reviews to predict ratings in practical situations. Therefore, *Task 2* can be more suitable for practical evaluations.

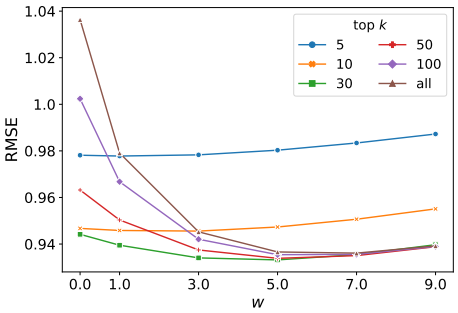

Figure 3: Effect of w and k. The y-axis represents RMSE, and x-axis represents w in Eq. 9. Each line represents k.

	Movies	Electronics	Clothing	Books
(A) *LF*	1.1002	1.4076	1.2061	1.0369
(B) *HIALF*	1.0644	1.3630	1.1739	1.0182
(C) *Simi-avg.*	1.0589	1.2015	**1.0789**	0.9332
(D) *Ours*	**1.0068**	**1.2001**	1.0800	**0.9261**
Ours vs *LF*	8.49%	14.74%	10.45%	10.68%
Ours vs *HIALF*	5.41%	11.95%	8.00%	9.04%

Table 4: Results of rating predictions (RMSE).

Results of intrinsic rating prediction in Task2: Table 5 shows the results of the intrinsic rating prediction. However, it is considered that the aspects of the pseudo-users are synthesized from many aspects of users. We consider that the RMSE calculated including the pseudo-user might contain noise. Thus, for a more personalized evaluation, we calculated RMSE without ratings of pseudo-users. In Table 5, the results show that our proposed model outperforms *HIALF* except for the Electronics dataset. This indicates that our model succeeds in separating historical influence and learning intrinsic preferences. Table 6 shows that our model has a better performance than *HIALF* in personalized evaluations. From these results, we consider that our model can predict the intrinsic ratings more accurately.

	Movies	Electronics	Clothing	Books
HIALF	1.0437	**1.2112**	1.0745	0.8971
Ours	**1.0302**	1.2118	**1.0673**	**0.8672**

Table 5: Results of predicting the first ratings. RMSE are reported in the table.

	Movies	Electronics	Clothing	Books
HIALF	1.0076	0.9222	(0.6886)	0.7388
Ours	**0.9959**	**0.8854**	(**0.6208**)	**0.7378**

Table 6: Results of predicting the first ratings **except for pseudo-users**. In the Clothing category, there is only one rating given by non-pseudo-user (reported in parentheses).

7 Conclusion

In this study, we propose an aspect-similarity-aware historical influence model for rating prediction. First, we perform preliminary experiments to validate our assumption that review aspects relate to the historical influence. From the analyses, we concluded that the sentence similarity relates to the historical influence. Thus, to model historical influence, we used textual information in user reviews, which previous models have ignored, for topic modeling. In the ordinay rating prediction task, we showed that the proposed approach achieved improvements over the previous historical-influence-aware models. Furthermore, to examine whether the proposed model can distinguish the intrinsic ratings of users from the historical influence, we conducted intrinsic rating prediction experiments. The results showed that our model has better performance than previous models.

Our method is limited to situations where the reviews of users are obtained before they provide ratings. For usual recommendation, it might not be practical. Thus, the method is not suitable for recommendations based on rating predictions. However, debiasing the historical influence to obtain intrinsic ratings can be considered as an application of our model. The intrinsic ratings would help subsequent users to make unbiased purchase decisions. In addition, recommender systems would be able to suggest acceptable items to users by using the intrinsic ratings. These improvements will provide better user experience of the e-commerce services.

References

Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2016. Understanding Effects of Personalized vs. Aggregate Ratings on User Preferences. In *Proceedings of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, IntRS '16, pages 14–21.

Rolph E. Anderson. 1973. Consumer Dissatisfaction: The Effect of Disconfirmed Expectancy on Perceived Product Performance. *Journal of Marketing Research*, 10(1):38–44.

Edward Loper Bird, Steven and Ewan Klein. 2009. Natural language processing with python.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural Attentional Rating Regression with Review-level Explanations. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1583–1592.

Guibing Guo, Jie Zhang, Daniel Thalmann, Anirban Basu, and Neil Yorke-Smith. 2014. From Ratings to Trust: An Empirical Study of Implicit Trust in Recommender Systems. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pages 248–253.

Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 507–517.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37.

Munan Li, Kenji Tei, and Yoshiaki Fukazawa. 2019a. An efficient co-Attention Neural Network for Social Recommendation. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, pages 34–42.

Pengfei Li, Hua Lu, Gang Zheng, Qian Zheng, Long Yang, and Gang Pan. 2019b. Exploiting Ratings, Reviews and Relationships for Item Recommendations in Topic Based Social Networks. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 995–1005.

Guang Ling, Michael R. Lyu, and Irwin King. 2014. Ratings Meet Reviews, a Combined Approach to Recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*, RecSys '14, pages 105–112.

Yiming Liu, Xuezhi Cao, and Yong Yu. 2016. Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 269–272.

Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the 7th ACM conference on Recommender systems*, RecSys '13, pages 165–172.

Subhabrata Mukherjee and Stephan Guennemann. 2019. GhostLink: Latent Network Inference for Influence-aware Recommendation. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 1310–1320.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, May.

Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-Boosted Latent Topics: Understanding Users and Items with Ratings and Reviews. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2640–2646, July.

Ting Wang, Dashun Wang, and Fei Wang. 2014. Quantifying Herding Effects in Crowd Wisdom. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1087–1096.

Le Wu, Peijie Sun, Yanjie Fu, Richang Hong, Xiting Wang, and Meng Wang. 2019a. A Neural Influence Diffusion Model for Social Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, pages 235–244.

Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Peng He, Paul Weng, Han Gao, and Guihai Chen. 2019b. Dual Graph Attention Networks for Deep Latent Representation of Multifaceted Social Effects in Recommender Systems. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 2091–2102.

Mao Ye, Xingjie Liu, and Wang-Chien Lee. 2012. Exploring Social Influence for Recommendation - A Generative Model Approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 671–680.

Xiaoying Zhang, Hong Xie, Junzhou Zhao, and John C. S. Lui. 2019. Understanding Assimilation-contrast Effects in Online Rating Systems: Modelling, Debiasing, and Applications. *ACM Trans. Inf. Syst.*, 38(1):2:1–2:25.

Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. 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*, WSDM '17, pages 425–434.