Aspect Mining with Rating Bias

Yitong Li[1], Chuan Shi[1,**], Huidong Zhao[1], Fuzhen Zhuang[2], and Bin Wu[1]

[1]Beijing Key Lab of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, China
{liyitong, shichuan, wubin}@bupt.edu.cn, zhaohuidong1121@foxmail.com
[2]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
zhuangfz@ics.ict.ac.cn

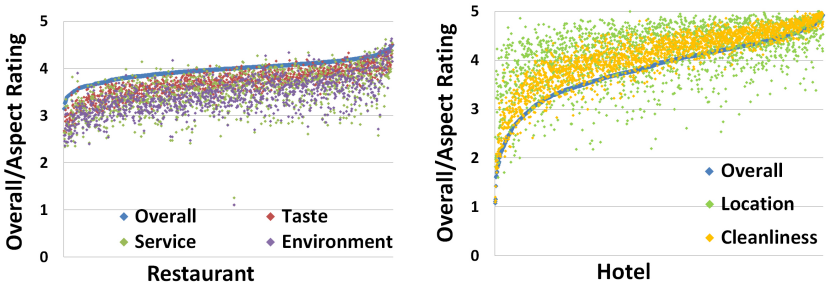
Abstract. Due to the personalized needs for specific aspect evaluation on product quality, these years have witnessed a boom of researches on aspect rating prediction, whose goal is to extract ad hoc aspects from online reviews and predict rating or opinion on each aspect. Most of the existing works on aspect rating prediction have a basic assumption that the overall rating is the average score of aspect ratings or the overall rating is very close to aspect ratings. However, after analyzing real datasets, we have an insightful observation: there is an obvious rating bias between overall rating and aspect ratings. Motivated by this observation, we study the problem of aspect mining with rating bias, and design a novel RAting-center model with BIas (RABI). Different from the widely used review-center models, RABI adopts the overall rating as the center of the probabilistic model, which generates reviews and topics. In addition, a novel aspect rating variable in RABI is designed to effectively integrate the rating bias priori information. Experiments on two real datasets (Dianping and TripAdvisor) validate that RABI significantly improves the prediction accuracy over existing state-of-the-art methods.

Keywords: aspect rating; rating prediction; rating bias; topic model

1 Introduction

With the rapid development of the Internet, the information which people can gain from the Internet grows exponentially. Nowadays, people are used to viewing online reviews before making decisions. For example, if a user wants to go out for dinner, he or she may look at the reviews of restaurants around on the Internet and choose one according to his or her taste. These reviews contain mainly overall ratings which evaluate restaurants from a general view. However, people may expect more subtle aspect ratings, such as the taste, environment, service, and so on. This problem has inspired the research on aspect-level opinion mining. The goal of the aspect-level opinion mining (i.e., aspect identification and aspect rating prediction) is to extract ad hoc aspects from online reviews and predict rating or opinion on each aspect.

** Corresponding author.



(a) Rating on Dianping (b) Rating on TripAdvisor

Fig. 1. Distributions of ratings on Dianping and TripAdvisor

Because of its great practical significance, there is a surge of researches on aspect identification and aspect rating prediction in recent years. Some works generate ratable aspects for reviews with whole overall ratings [7] or scarce overall ratings [6], and some works consider to integrate external knowledge [9]. Most of the existing works predict aspect ratings with the help of overall ratings, and they all have a basic assumption. That is, the overall rating is the average score of aspect ratings or the overall rating is close to aspect ratings.

However, the analysis on real datasets shows an insightful phenomenon: there is an obvious and systemic rating bias between overall ratings and aspect ratings. Fig. 1 illustrates the rating distributions on two real datasets: Dianping[1] (a well-known social media platform in China, which contains the information and reviews of restaurant, hotel, entertainment, movie, etc) and TripAdvisor[2] (a widely used dataset in this field, which is a social media platform about travel, hotel, scenic spot, etc). The datasets we use are the restaurant data in Dianping and the hotel data in TripAdvisor. Note that the overall ratings of restaurants/hotels are sorted in an ascending order in Fig. 1. We can find that the overall ratings in TripAdvisor are obviously lower than two aspect ratings, while the overall ratings in Dianping are significantly larger than aspect ratings. The interesting observation implies that the previous aspect rating prediction approaches may achieve poor performance, if ignoring the rating bias between overall ratings and aspect ratings.

Motivated by the observed rating bias, we try to study the problem of aspect mining with rating bias. That is, the goal is to decompose the reviews into different aspects and predict the rating of different aspects on each entity, with the help of the overall rating and the rating bias priori information. However, aspect mining with rating bias may face two challenges. First, the rating process of users may conform to some behaviour patterns, which determine the dependency relationship among the variables in the topic model. Most of the existing works on aspect rating prediction are based on probabilistic graphical model.

[1] http://www.dianping.com/
[2] http://www.tripadvisor.com/

Inspired by the word generation process, these works usually consider ratings are finally generated by reviews, topics or aspects. However, does it really comply to user behaviour? We have a different view. We believe that users form an intuitive impression (good or bad) as soon as they experienced the product, which is reflected by rating. Only after the impression (rating) is formed will the user write a review (or words) to express his/her feeling. So we think the previous models may not conform to user behaviour properly, and thus we need to mine the authentic rating behaviour of users. Second, how to effectively utilize the rating bias information? As we mentioned above, there is an obvious bias between overall rating and aspect ratings. The rating bias may cause the inaccuracy of aspect rating prediction, and influence the results tremendously. Luo et al. [6] have discovered the rating bias, but nobody has considered it in the model until now. So how to use the rating bias priori information properly to improve the prediction accuracy is also a challenge.

To solve the challenges mentioned above, we design a novel RAting-center model with BIas (RABI). Different from traditional rating generating process [7, 6, 9], RABI considers rating as the center of the model, which generates the reviews and topics. This idea stems from users' real experiences. When users decide to write a review, they usually have intuitional opinions (i.e., overall ratings) on the products, and then they will use proper phrases to represent their opinions. In addition, RABI introduces a novel latent aspect rating variable which can effectively learn the correlation of the overall rating, aspects, and rating bias. Experimental results on two real datasets (i.e., Dianping and TripAdvisor) validate the effectiveness of RABI on both Chinese and English reviews, compared to existing state-of-the-art methods. The results also show that RABI can accurately decompose the reviews into different aspects.

Our contributions are summarized as follows:

- We first analyze the rating bias between overall rating and aspect ratings in real data, and put forward the problem of aspect mining with rating bias.
- We propose a novel RABI model for aspect mining with rating bias. Different from existing models, RABI considers rating as the center of the model, which simulates the generation of the review better. In addition, an aspect rating variable is proposed to effectively utilize the rating bias information.
- Experiments on real datasets have shown the effectiveness of our algorithm over existing state-of-the-art methods.

2 Data Analysis

In order to show the rating bias phenomenon, we analyze two real datasets. The first dataset is crawled from Dianping website, a well-known social media platform in China, which provides a review platform for businesses and entertainments. In Dianping website, a user can give a review to a business after enjoying a service in this business. Besides an overall rating, the review information includes Chinese comments and three aspect ratings on Taste, Service, and Environment, respectively. In addition, we also employ the widely used TripAdvisor dataset

Table 1. Statistics of the datasets

Datasets	#Products	#Reviews	#Phrases	Avg. Overall Rating
Dianping	1,097	216,291	696,608	3.97
TripAdvisor	1,850	197,970	2,571,902	3.81

Table 2. Rating bias on each aspect on both datasets

Dataset	Category	Avg. Rating	Rating Bias
Dianping	*Overall*	3.97	
	Taste	3.69	+0.28
	Service	3.48	+0.48
	Environment	3.43	+0.54
TripAdvisor	*Overall*	3.81	
	Value	3.80	+0.01
	Room	3.82	-0.01
	Location	4.14	-0.33
	Cleanliness	4.07	-0.26
	Front Desk/Staff	3.96	-0.15
	Service	3.92	-0.11
	Business	3.59	+0.22

[10]. Accompanying with English comments, reviews in this dataset are not only associated with overall ratings, but also with ground truth aspect ratings on 7 aspects: Value, Room, Location, Cleanliness, Front desk/staff, Service, and Business. All the ratings in the datasets are in the range from 1 to 5. The statistic information of these datasets is shown in Table 1.

We first intuitively show the distributions of overall and aspect ratings on these two datasets in Fig. 1. Note that, we only show the distributions of some aspect ratings due to the space limitation. Moreover, we sort products according to their overall ratings for clarity. From Fig. 1, we can find that there are obvious rating biases between overall rating and aspect rating on both datasets. In Dianping dataset, the overall rating is far above the aspect ratings in all three aspects, while the overall rating is smaller than two aspect ratings in TripAdvisor.

Furthermore, we calculate the rating bias on each aspect on both datasets. The calculating process can be seen in Eq.(1) and the results are listed in Table 2. The rating biases in Dianping are huge on most aspects, especially +0.48 for Service and +0.54 for Environment, which are pretty huge values. So the rating biases in Dianping should be well considered. The rating biases in TripAdvisor are small on some aspects (e.g., +0.01 for Value and -0.01 for Room), but huge on other aspects (e.g., -0.33 for Location and -0.26 for Cleanliness). Although the rating biases in TripAdvisor are not as much as those in Dianping, they all truly exist. The interesting observation implies that the previous aspect rating prediction approaches may achieve poor performance, if ignoring the rating bias. As shown in Table 2, the rating biases are different in different datasets

and aspects, which can influence the results to varying degrees and cause the inaccuracy of aspect rating prediction. So the proper consideration of the rating bias can improve the prediction accuracy.

3 Preliminary Notations and Problem Definition

In this section, we first introduce the notations and concepts used in this paper, and then formally propose the problem of aspect mining with rating bias.

Entity: An entity e indicates a product which belongs to the product set E (e.g., a restaurant in Dianping dataset or a hotel in TripAdvisor dataset). N_e indicates the number of entities in E.

Review: A review d is the user's opinion about the entity e. An entity e can have many reviews from different users. A review consists of the text content, the overall rating and many aspect ratings. There are N_d reviews in total.

Phrase: A phrase $f = (h, m)$ consists of a pair of words, which are extracted from the review's text content. h denotes the head term, and m is the modifier term which modifies h. A review d contains several phrases f.

Head term: The head term h is used to describe the aspect information. It decides which aspect the phrase f is expressing. For instance, "attitude" is a head term, and it belongs to the aspect "Service".

Modifier term: The modifier term m is used to describe the sentiment information. It is used to describe the aspect, which is decided by h, is good or bad. For instance, for the head term "attitude", "cold" or "passionate" may be used as the modifier term.

Overall rating: An overall rating r of a review d is a numerical rating, which indicates the user's overall sentiment tendency on the entity e. The number of the values of rating is N_r and it is usually 5, which means the values of rating r are from 1 to 5.

Aspect: An aspect A_i is a specific side of the entity e, e.g., the taste of the restaurant. It is a set of many similar characteristic of the entity e. N_A indicates the number of aspects.

Aspect rating: An aspect rating r_{A_i} is a numerical rating, which indicates the user's sentiment tendency on the aspect A_i of the entity e, and is also from 1 to 5. And a review d has N_A aspect ratings, which corresponds to N_A aspect.

Rating bias: The rating bias is the gap between the average of overall ratings and the average of aspect ratings. There are N_A biases on N_A aspects, and they are in connection with the current aspect A_i. The rating bias b_{A_i} on aspect A_i can be calculated as follows:

$$b_{A_i} = \frac{\sum_d r}{N_d} - \frac{\sum_d r_{A_i}}{N_d}. \tag{1}$$

Aspect mining with rating bias: The problem of *aspect mining with rating bias* is to predict the rating on each aspect with the rating bias prior information. Specifically, given a set of reviews $D = \{d_1, d_2, \cdots, d_{N_d}\}$ about entities $E = \{e_1, e_2, \cdots, e_{N_e}\}$, we know that each review $d_i \in D$ contains text

content (Chinese or English) and overall rating r on an entity $e_j \in E$, as well as the rating bias b_{A_i} between the overall rating and the aspect rating on N_A aspects for all reviews. The goal is to decompose the phrases f, which are extracted from texts in D, into N_A aspects $\{A_1, A_2, \cdots, A_{N_A}\}$, and rate the aspects of each entity e with $\{r_{A_1}, r_{A_2}, \cdots, r_{A_{N_A}}\}$.

In fact, our goal includes two sub-tasks. (1) The first sub-task is aspect identification, which is to correctly identify the aspect label A_i given phrase f. (2) The second sub-task is aspect rating prediction, which is to predict the aspect rating r_{A_i} given the entity e and aspect A_i.

The problem of aspect mining with rating bias is very important in real applications. The problem is also the base of many tasks, such as overall rating prediction and aspect-level product recommendation. Compared to overall ratings, the aspect ratings are always missing and more unreliable. The aspect rating prediction is an effective way to repair the missing ratings and correct the unreliable ratings. However, the existent rating bias may make current methods on aspect rating prediction not effective anymore, so it is desired to consider rating bias for aspect rating prediction. Please note that the rating bias is known in our problem setting. Moreover, the rating bias can be easily obtained through limited reliable aspect ratings or a small quantity of manual labeling in real applications. So we can use the information of rating bias to correct the aspect rating prediction.

4 Rating-Center Model with Bias

The simplest way to handle rating bias is to subtract rating bias from the rating prediction results of existing models. However, it does not consider the correlations of ratings, aspects, and rating bias, so it may result in poor performances. In this section, we propose a novel RABI to handle the problem of the existent rating bias. Furthermore, we derive an iterative optimization solution with the EM algorithm.

4.1 Model Description

Existing models on aspect rating prediction usually consider reviews as the center to generate ratings and topics [6, 9, 7]. However, it does not conform to the authentic rating behaviour of users. In daily life, we form an intuitive impression as soon as we experienced a product. Only after we form an intuitive opinion (like or dislike, quantitively represented by a rating) on a product, will we write a review to express our opinion. In addition, our opinion may involve multiple aspects of the product, such as taste, service and environment. So in the generative process of a product review, we will choose proper head terms to represent the aspect we want to express, and proper modifier terms to express sentiments on corresponding head terms. Finally, we organize these terms and other words to form a review. Therefore, we believe it is more reasonable to consider rating (overall rating) as the center to generate topics and reviews, which conforms

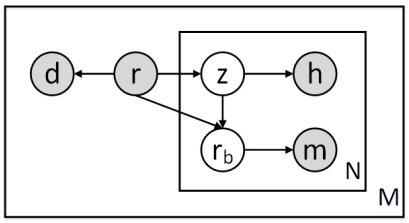


Fig. 2. Graphical model of RABI

to the authentic rating behaviour of users. Following this idea, we design the probabilistic model of RABI, shown in Fig. 2.

In Fig. 2, d indicates the reviews, r indicates the overall rating, h indicates the head term and m indicates the modifier term. These four variables are represented as the shaded circles, which means these four variables are observable. z indicates the aspect A_i. In order to keep consistent with the topic model, the aspect A_i is expressed as the topic z. And r_b indicates aspect rating, which will be introduced in the following. These two variables are represented as the open circles, which means these two variables are latent variables. Furthermore, N indicates the number of phrases in a review. And M indicates the number of reviews, which is equal to N_d.

To utilize the rating bias information effectively, we bring in a new latent aspect rating variable r_b. The modifier term m is used to modify the head term h to express the opinion (like or dislike) on aspect A_i (represented with z in the model), so m is actually influenced by the corresponding aspect rating r_{A_i}. As we mentioned above, there is an obvious rating bias between overall rating and aspect ratings. This observation causes that we cannot use the overall rating r to influence the modifier term m directly. So we bring in a new variable r_b between r and m to eliminate the influence of rating bias. r_b indicates an unknown aspect rating, so it is a latent variable. For a certain aspect A_i, the value of r_b is set as the overall rating r minus the rating bias b_{A_i}. Note that r_b can take N_r values in A_i, since the variable r can take N_r values. By bringing in the latent variable r_b, the association between r and m is modeled more reasonably in RABI.

According to the RABI model shown in Fig. 2, as the origin of the model, the overall rating r generates the review d and the latent topic z. The latent aspect rating r_b depends on the topic z and the overall rating r. And the head term h and the modifier term m are influenced by the topic z and the aspect rating r_b, respectively. So the joint probability over all variables is as follows:

$$p(h, m, r, d, z, r_b) = p(m|r_b)p(r_b|r, z)p(h|z)p(z|r)p(d|r)p(r). \tag{2}$$

All the parameters can be iteratively calculated using the EM algorithm [4], which is a common method to solve the problem with latent variable. The detail derivation is given in next section.

4.2 EM Solution

In the E-step, we need to maximize the lower bound function $\mathcal{L}_0$ (i.e., Jensens inequality [2]),

$$\mathcal{L}_0 = \sum_{z,r_b} q(z, r_b) \log\{\frac{p(h, m, r, d, z, r_b|\Lambda)}{q(z, r_b)}\}. \tag{3}$$

Here, as usual, $q(z, r_b)$ is set as follows:

$$q(z, r_b) = p(z, r_b|h, m, r, d; \Lambda^{old}). \tag{4}$$

Then we simplify Eq.(3), we can get

$$\begin{aligned}\mathcal{L}_0 &= \sum_{z,r_b} q(z, r_b) \log\{\frac{p(h, m, r, d, z, r_b|\Lambda)}{q(z, r_b)}\} \\ &= \underbrace{\sum_{z,r_b} q(z, r_b) \log p(h, m, r, d, z, r_b|\Lambda)}_{\mathcal{L}} - \underbrace{\sum_{z,r_b} q(z, r_b) \log q(z, r_b)}_{const} \\ &= \mathcal{L} - const.\end{aligned} \tag{5}$$

So the second part is a *const*, which can be ignored. Then we ignore the *const*, and only consider the $\mathcal{L}$.

The function for the posterior probabilities of the latent variables is as follows:

$$\mathcal{L} = \sum_{h,m,r,d,z,r_b} n(h, m, r, d) q(z, r_b) \log p(h, m, r, d, z, r_b|\Lambda), \tag{6}$$

where Λ includes all parameters, i.e., $p(m|r_b)$, $p(r_b|r, z)$, $p(h|z)$, $p(z|r)$, $p(d|r)$ and $p(r)$, which are mentioned in Eq.(2). Besides, $n(h, m, r, d)$ is the number of co-occurrences of h, m, r and d.

The function $q(z, r_b)$ and $p(h, m, r, d, z, r_b|\Lambda)$ in Eq.(3) are expanded as follows:

$$q(z,r_b) = p(z, r_b|h, m, r, d; \Lambda^{old}) = \frac{p(m|r_b)p(r_b|r, z)p(h|z)p(z|r)p(d|r)p(r)}{\sum_{z,r_b} p(m|r_b)p(r_b|r, z)p(h|z)p(z|r)p(d|r)p(r)}, \tag{7}$$

$$p(h, m, r, d, z, r_b|\Lambda) = p(m|r_b)p(r_b|r, z)p(h|z)p(z|r)p(d|r)p(r). \tag{8}$$

In the M-step, the Lagrangian Multiplier method is used to maximize $\mathcal{L}$ and calculate the parameters.

For $p(m|r_b)$, there is a basic constraint as follows:

$$\sum_{m} p(m|r_b) = 1. \tag{9}$$

Applying the Lagrangian Multiplier method, we can get a function for $p(m|r_b)$ as follows:

$$\frac{\partial[\mathcal{L}_{[p(m|r_b)]} + \lambda(\sum_m p(m|r_b) - 1)]}{\partial p(m|r_b)} = 0. \tag{10}$$

After calculation, we have

$$p(m|r_b) \propto n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old}). \tag{11}$$

Then the update function for $p(m|r_b)$ is as follows:

$$p(m|r_b) = \frac{\sum\limits_{h,r,d,z} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h,m',r,d,z} n(h, m', r, d)p(z, r_b|h, m', r, d; \Lambda^{old})}. \tag{12}$$

Similarly, the update functions for other parameters are as follows:

$$p(r_b|r, z) = \frac{\sum\limits_{h,m,d} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h,m,d,r_b{}'} n(h, m, r, d)p(z, r_b{}'|h, m, r, d; \Lambda^{old})}, \tag{13}$$

$$p(h|z) = \frac{\sum\limits_{m,r,d,r_b} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h',m,r,d,r_b} n(h', m, r, d)p(z, r_b|h', m, r, d; \Lambda^{old})}, \tag{14}$$

$$p(z|r) = \frac{\sum\limits_{h,m,d,r_b} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h,m,d,z',r_b} n(h, m, r, d)p(z', r_b|h, m, r, d; \Lambda^{old})}, \tag{15}$$

$$p(d|r) = \frac{\sum\limits_{h,m,z,r_b} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h,m,d',z,r_b} n(h, m, r, d')p(z, r_b|h, m, r, d'; \Lambda^{old})}, \tag{16}$$

$$p(r) = \frac{\sum\limits_{h,m,d,z,r_b} n(h, m, r, d)p(z, r_b|h, m, r, d; \Lambda^{old})}{\sum\limits_{h,m,r',d,z,r_b} n(h, m, r', d)p(z, r_b|h, m, r', d; \Lambda^{old})}. \tag{17}$$

Through these functions above, we can iteratively calculate the parameters until the model has converged.

4.3 Aspect Rating Prior

To verify our model's effectiveness, we need to compare the predicted aspect ratings with the real aspect ratings. So the aspects should correspond to the real aspects which are set by the e-commerce review sites. To make the predicted aspects similar to the real aspects, we need to assign some seed words to each

aspect. For instance, the aspect "Taste" may include a few prior words, such as "taste" and "flavor".

In our model, we inject the prior knowledge for the aspect z. The function is as follows:

$$p(h|z) = \frac{\sum\limits_{m,r,d,r_b} n(h,m,r,d)p(z,r_b|h,m,r,d;\Lambda^{old}) + \tau(h,z)}{\sum\limits_{h',m,r,d,r_b} n(h',m,r,d)p(z,r_b|h',m,r,d;\Lambda^{old}) + \sum\limits_{h'} \tau(h',z)}, \tag{18}$$

where $\tau(h,z)$ indicates the prior knowledge of the prior words. Only when there is a relationship between the head term h and the topic z, in other words, h belongs to z, does $\tau(h,z)$ have a value δ, otherwise 0.

Note that, in the real applications, we can set aspects manually or generate aspects by the model directly. Moreover, manual aspect setting usually has better performances.

4.4 Aspect Identification and Aspect Rating Prediction

We can get $p(z,r_b|h,m)$ from the model by the following function,

$$\begin{aligned} p(z,r_b|h,m) &= \frac{\sum_{r,d} p(h,m,r,d,z,r_b)}{\sum_{r,d,z,r_b} p(h,m,r,d,z,r_b)} \\ &= \frac{\sum_{r,d} p(m|r_b)p(r_b|r,z)p(h|z)p(z|r)p(d|r)p(r)}{\sum_{r,d,z,r_b} p(m|r_b)p(r_b|r,z)p(h|z)p(z|r)p(d|r)p(r)}. \end{aligned} \tag{19}$$

The goal of aspect identification is to find the mapping function $\mathcal{G}$ that correctly assigns the aspect label for given phrase f.

$$\mathcal{G}(f=(h,m)) = \arg\max_z \sum_{r_b} p(z,r_b|h,m). \tag{20}$$

The goal of aspect rating prediction is to predict the aspect rating r_{A_i} of the entity e given all the phrases f from all reviews and aspect $A_i(z)$. The aspect rating function is as follows:

$$r_{e,A_i} = \frac{\sum_{(h,m)\in\text{all reviews of } e} \sum_{r_b} r_b \cdot p(z,r_b|h,m)}{\sum_{(h,m)\in\text{all reviews of } e} \sum_{r_b} p(z,r_b|h,m)}, \tag{21}$$

where r_{e,A_i} indicates the aspect rating on the aspect A_i of the entity e.

In this way, RABI learns the joint probability distribution of phrases, aspects and ratings, and predicts aspect ratings with bias.

5 Evaluation

In this section, we introduce experimental preparation, evaluation metric and baselines. Then we conduct extensive experiments to evaluate the effectiveness of RABI on two real datasets.

Table 3. Prior words for aspect prior

Dataset	Category	Prior Words
Dianping	Taste	taste, flavor, dish, dishes
	Service	serving, attitude, waitress, service
	Environment	environment, location, room, decoration
TripAdvisor	Value	value, price, quality, worth
	Room	room, suite, view, bed
	Location	location, traffic, place, area
	Cleanliness	clean, dirty, maintain, smell
	Front Desk/Staff	staff, check, help, reservation
	Service	service, food, breakfast, buffet
	Business	business, center, computer, internet

5.1 Experimental Preparation

Experiments are conducted on two real datasets (i.e., Dianping and TripAdvisor), which are introduced in Section 2. The preprocessing of TripAdvisor is similar to that in [6]. But the preprocessing of Dianping is slightly different. Since Dianping is a Chinese website, the Word Segmenter[3] and the rules from [8] are adopted for preprocessing. To inject the prior knowledge for the aspect, we select some words as prior for each aspect, and Table 3 lists some of the prior words (not all of the prior words due to the space limitation). For better understanding, we translate the Chinese words in Dianping into English.

Besides, all of the initial parameters ($p(m|r_b)$, $p(r_b|r,z)$, $p(h|z)$, $p(z|r)$, $p(d|r)$ and $p(r)$ in Eq.(2)) are assigned uniformly and randomly. δ in the Section 4.3 is set as 1 after some preliminary tests. The number of aspects or topics K is set as 3 for Dianping and 7 for TripAdvisor. The experiments are done on different-size of datasets (i.e., 25%, 50%, 75%, and 100% of review data) from Dianping and TripAdvisor, respectively. The maximum number of iterations is set as 500.

5.2 Evaluation Metric

RMSE (Root Mean Square Error) is one of the most common metrics for rating prediction. RMSE can measure the difference between the real values and the predicted values. For every entity e, we have the real aspect rating vector r_{e,A_i} and the predicted aspect rating vector $\hat{r}_{e,A_i}$. The function of RMSE is as follows:

$$RMSE = \sqrt{\frac{\sum_{e=0}^{N_e} \sum_{A_i=0}^{N_A} (\hat{r}_{e,A_i} - r_{e,A_i})^2}{N_e * N_A}} \tag{22}$$

Smaller value of RMSE indicates a stronger predictor, which means the real values and the predicted values are nearer.

Besides, we use Pearson Correlation Coefficient ρ [10] to measure the relative ordering of products based on the predicted aspect rating and the real aspect

[3] http://nlp.stanford.edu/software/segmenter.shtml

rating. The correlation is stronger when the absolute value of ρ is closer to 1, and weaker when the absolute value of ρ is closer to 0. The function is as follows:

$$\rho = \frac{N \sum \hat{r}_{e,A_i} r_{e,A_i} - \sum \hat{r}_{e,A_i} \sum r_{e,A_i}}{\sqrt{N \sum (\hat{r}_{e,A_i})^2 - (\sum \hat{r}_{e,A_i})^2} \sqrt{N \sum (r_{e,A_i})^2 - (\sum r_{e,A_i})^2}}, \tag{23}$$

where N indicates the total amount, which is $N_e * N_A$.

5.3 Baseline Methods

We compare the proposed model with three representative methods and one variation of RABI. Since all of these baselines do not consider the rating bias, we adjust the results of these baselines through subtracting the rating bias for fair comparison. The adjusted method is marked with "*" to distinguish from the original method.

- QPLSA/QPLSA* [7] uses quad-tuples information to build a model based on PLSA framework. The model not only can generate fine-granularity aspects of products, but also capture the relationship between words and ratings.
- GRAOS/GRAOS* [6] is a semi-supervised model based on LDA framework. It also uses the quad-tuples information to capture the relationship between words and ratings. The model considers the rating distribution as a Gaussian distribution.
- SATM/SATM* [9] is a sentiment-aligned model based on LDA framework. The model uses two kinds of external knowledge: productlevel overall rating distribution and wordlevel sentiment lexicon.
- RA/RA* is a simplified model which removes the latent aspect rating variable r_b from our model RABI. It only considers the rating-center assumption. Through comparing RA* and RABI, we can testify the importance of the good mechanism to utilize rating bias information.

5.4 Results Evaluation

We firstly validate the effectiveness of aspect identification of RABI through a case study, and then compare the results of different methods on the accuracy of aspect rating prediction with two criteria mentioned above.

Aspect Identification RABI extracts a set of rated phrases to describe the product for each aspect. We list the top 20 automatically mined phrases for each aspect, from which we select several meaningful phrases to be shown in Table 4. The phrases are ranked by their ratings for every aspect.

Generally, the extracted phrases properly describe the corresponding aspects and accurately embody the opinion in both English and Chinese reviews. On one hand, the head terms can indicate the aspects well, such as "attitude" for service, "fitment" for environment, "setting" for room, and "area" for location.

Table 4. Representative phrases for different aspects on two datasets

Datasets	Aspects	Representative Phrases(Ratings)
Dianping	Taste	amazing mouthfeel(4.71), first-rate taste(4.58), common taste(2.75), so-so flavor(1.77)
	Service	smart waiter(4.51), passive service(3.51), slow serving(2.51), cold attitude(1.67)
	Environment	great location(4.45), sumptuous fitment(4.26), common environment(2.88), small room(2.45)
TripAdvisor	Value	perfect price(4.81), standard charge(4.65), delightful priceline(4.05), astronomical deal(1.59)
	Room	greatest setting(4.81), cool room(4.19), beautiful decor(4.18), worst setting(1.57)
	Location	wonderful location(4.90), central location(4.63), nice place(4.13), remote area(1.27)
	Cleanliness	normal maintained(4.61), standard cleanliness(4.38), well homey(4.34), dirty housekeeping(1.26)
	Front Desk/Staff	hospitable staff(4.95), great staff(4.71), friendly hotel(4.65), so-so staff(1.82)
	Service	super singer(4.71), great wine(4.50), valuable amenities(4.27), worst experience(1.23)
	Business	best wifi(4.63), common websites(4.22), nice desktop(4.17), standard business(3.52)

When a user sees the head term, he can understand which aspect is talked about. On the other hand, a positive modifier term indicates a positive attitude and is likely to obtain a higher rating, and a negative modifier term indicates a negative attitude and is likely to obtain a lower rating. For example, in the Service aspect of Dianning, the phrase "cold attitude" is rated as 1.67 because "cold" is a negative modifier term, while the phrase "smart waiter" has a score of 4.51 because "smart" is a positive modifier term. In addition, the phrases and their ratings are also able to reflect the different rating styles in Chinese and English. That is, users tend to give relativly lower ratings in Chinese reviews. The distribution of the predicted ratings on phrases also conforms to that of aspect-level ratings on these two datasets in Table 2. It also confirms the effectiveness of RABI on Chinese and English datasets.

Accuracy Experiment Then we validate the performances of different methods through comparing predicted aspect ratings with real aspect ratings using the RMSE criterion by Eq.(22).

From the results shown in Table 5, we can clearly find that the integration of the rating bias information can significantly improve the prediction accuracy for all methods (e.g., QPLSA* has better performances than QPLSA), and RABI always performs best on both datasets. The improvement is particularly obvious for Dianping, because this dataset has large rating biases. Although the rating bias is small in TripAdvisor, the methods considering rating bias all achieve

Table 5. RMSE performances of different methods on two datasets

	Dianping				**TripAdvisor**			
	25%	**50%**	**75%**	**100%**	**25%**	**50%**	**75%**	**100%**
QPLSA	0.5816	0.5799	0.5714	0.5635	0.6374	0.6248	0.6129	0.6119
QPLSA*	0.3656	0.3584	0.3554	0.3435	0.6262	0.6180	0.6125	0.6071
GRAOS	0.4751	0.4668	0.4624	0.4500	0.6056	0.6072	0.6011	0.5968
GRAOS*	0.4228	0.4152	0.4155	0.4136	0.5790	0.5724	0.5691	0.5668
SATM	0.5804	0.5767	0.5639	0.5594	0.5587	0.5502	0.5406	0.5300
SATM*	0.3979	0.3890	0.3816	0.3738	0.5419	0.5322	0.5310	0.5188
RA	0.5789	0.5601	0.5511	0.5451	0.6081	0.5935	0.5826	0.5784
RA*	0.3471	0.3404	0.3312	0.3267	0.5785	0.5612	0.5599	0.5471
RABI	**0.3248**	**0.3162**	**0.3047**	**0.2919**	**0.5346**	**0.5267**	**0.5204**	**0.5089**

Table 6. Pearson correlation coefficient of different methods on two datasets

	Dianping				**TripAdvisor**			
	25%	**50%**	**75%**	**100%**	**25%**	**50%**	**75%**	**100%**
QPLSA	0.5792	0.5809	0.5836	0.5985	0.3167	0.3451	0.3508	0.3827
GRAOS	0.1281	0.1280	0.1328	0.1376	0.3238	0.3407	0.3463	0.3569
SATM	0.3522	0.3605	0.3742	0.3906	0.3315	0.3521	0.3621	0.3679
RA	0.5248	0.5330	0.5430	0.5494	0.4065	0.4167	0.4291	0.4377
RABI	**0.6059**	**0.6137**	**0.6174**	**0.6211**	**0.5328**	**0.5522**	**0.5597**	**0.5657**

better performances than original methods. It illustrates that it is necessary to consider the rating bias for aspect rating prediction.

Besides, the rating-center model (i.e., RA) also achieves good performances among four baselines, which confirms the correctness of the rating-center assumption. Compared to simply subtracting the rating bias in four baselines, the best performances of RABI imply that the good mechanism to utilize rating bias information is also necessary. We think the rating-center and the latent aspect rating variable contribute to the good performances of RABI.

In addition, with the increment of review data, the accuracy of RABI increases steadily and slowly, which reflects that RABI is a steady method.

Relative Order Experiment Furthermore, we verify the ability of different methods to maintain the relative order among products with the Pearson Correlation Coefficient ρ. The results are shown in Table 6. Note that the rating bias has slight effect on the order of products, so we only display the results of original methods and ignore the adjusted methods. We can see that RABI obtains much higher ρ than other methods in all datasets. It once again shows that RABI is more effective to model the correlations between aspects and ratings, and thus better maintains aspect ranking orders compared to other methods. The results also imply that RABI is very promising for aspect-level recommender system, since it can generate very similar product order to the real order.

6 Related Work

In recent years, sentiment analysis on reviews becomes a research hotspot. Reviews focus on the products in each aspect, so sentiment analysis on reviews usually involves aspect. This situation leads to the aspect rating prediction. Aspect rating prediction usually contains two subtasks, aspect identification and aspect rating prediction.

Topic model is widely used to solve aspect identification. It mainly contains LSI [3], PLSA [5] and LDA [1]. Xu et al. [12] centered on implicit feature identification in Chinese product reviews via LDA and SVM. An AEP-based Latent Dirichlet Allocation (AEP-LDA) [13] model was also proposed to extract product and service aspect words automatically from reviews. Fu et al. [11] proposed an approach to automatically discover the aspects discussed in Chinese social reviews and classified the polarity of the associated sentiment by HowNet lexicon. Our model RABI is designed based on the PLSA framework.

To solve aspect identification and aspect rating prediction simultaneously, many researches adopted the topic-sentiment mixture models. QPLSA [7] adopted the quad-tuples, which consist of head, modifier, rating and entity. It can generate fine-granularity aspects and capture the correlations between words and ratings. SATM [9] used external knowledge, product-level overall rating distribution and word-level sentiment lexicon, to extract the product aspects and predict aspect ratings simultaneously. Luo et al. [6] proposed a model based on LDA to predict aspect ratings and overall ratings for unrated reviews and made two assumptions for the rating distribution. However, all of these works did not consider the existing rating bias, which is firstly studied in this paper.

7 Conclusion

Aspect rating prediction for reviews is a hot research issue nowadays. Most of researches base on such a basic assumption, the overall rating is the average score of aspect ratings or the overall rating is close to aspect ratings. However in the real world, there may be rating biases between overall rating and aspect ratings, and existing works did not consider these rating biases.

In this paper, we study the problem of aspect mining with rating bias and propose a novel probabilistic model RABI based on PLSA framework. The RABI model makes rating as the center to generate ratings and topics, and introduces a latent aspect rating variable to integrate the rating bias information. Experiments on two real datasets validate the effectiveness of RABI. In the future, we can import the Dirichlet prior and redesign our model based on LDA framework. The effectiveness will be enhanced further.

8 Acknowledgments

This work is supported in part by the National Key Basic Research and Department (973) Program of China (No. 2013CB329606), and the National Natural

Science Foundation of China (No. 61375058, 61473273), and the Co-construction Project of Beijing Municipal Commission of Education, and the CCF-Tencent Open Fund, and 2015 Microsoft Research Asia Collaborative Research Program.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, January 2003.
2. D. Chandler. Introduction to modern statistical mechanics. *Physics Today*, 1:288, September 1987.
3. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, September 1990.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39:1–38, 1977.
5. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, Berkeley, California, August 1999. Association for Computing Machinery.
6. W. Luo, F. Zhuang, X. Cheng, Q. He, and Z. Shi. Ratable aspects over sentiments: Predicting ratings for unrated reviews. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 380–389, Shenzhen, China, December 2014. Institute of Electrical and Electronics Engineers.
7. W. Luo, F. Zhuang, Q. He, and Z. Shi. Quad-tuple plsa: incorporating entity and its rating in aspect identification. In *Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference*, pages 392–404. Springer Berlin Heidelberg, Kuala Lumpur, Malaysia, May 2012.
8. S. Moghaddam and M. Ester. On the design of lda models for aspect-based opinion mining. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 803–812, Sheraton, Maui Hawaii, October 2012. Association for Computing Machinery.
9. H. Wang and M. Ester. A sentiment-aligned topic model for product aspect rating prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Doha, Qatar, October 2014. The Association for Computational Linguistics.
10. H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792, Washington DC, DC, July 2010. Association for Computing Machinery.
11. F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang. Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon. *Knowledge-Based Systems*, 37:186–195, 2013.
12. H. Xu, F. Zhang, and W. Wang. Implicit feature identification in chinese reviews using explicit topic mining model. *Knowledge-Based Systems*, 76:166–175, March 2015.
13. X. Zheng, Z. Lin, X. Wang, K.-J. Lin, and M. Song. Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. *Knowledge-Based Systems*, 61:29–47, 2014.