Aspect and Ratings Inference with Aspect Ratings: Supervised Generative Models for Mining Hotel Reviews

Wei Xue(✉), Tao Li, and Naphtali Rishe

Computer Science Department, Florida International University,
11200 Southwest 8th Street, Miami, FL33199, USA
{wxue004,taoli,rishe}@cs.fiu.edu
http://www.cis.fiu.edu

Abstract. Today, a large volume of hotel reviews is available on many websites, such as TripAdvisor (http://www.tripadvisor.com) and Orbitz (http://www.orbitz.com). A typical review contains an overall rating and several aspect ratings along with text. The rating is perceived as an abstraction of reviewers' satisfaction in terms of points. Although the amount of reviews having aspect ratings is growing, there are plenty of reviews including only an overall rating. Extracting aspect-specific opinions hidden in these reviews can help users quickly digest them without actually reading through them. The task mainly consists of two parts: aspect identification and rating inference. Most existing studies cannot utilize aspect ratings which are becoming abundant in the last few years. In this paper, we propose two topic models which explicitly model aspect ratings as observed variables to improve the performance of aspect rating inference over unrated reviews. Specifically, we consider sentiment distributions in the aspect level, which generate sentiment words and aspect ratings. The experiment results show our approaches outperform other existing methods on the data set crawled from TripAdvisor.

Keywords: Sentiment analysis · Information retrieval · Topic model

1 Introduction

The trend that people browse hotel reviews on websites before booking encourages researchers to focus on the analysis of the social media data. Users write down their own experience, and rate hotels with an overall score and/or along with several scores on aspects predefined by websites such as `room`, `service`, and `location`. Overall ratings express a general impression of reviewers which is more abstract than text, but they also hide aspect-specific sentiments. To this end, overall ratings are not informative enough. Although more and more reviews with aspect ratings are available on-line, there is a lot of reviews associated with only an overall rating. Therefore identifying aspect and learning more informative aspect ratings is an attractive topic in opinion mining, which helps users gain more details of each aspect.

J. Wang et al. (Eds.): WISE 2015, Part II, LNCS 9419, pp. 17–31, 2015.
DOI: 10.1007/978-3-319-26187-4_2

Many approaches have been proposed towards simultaneous aspect identification and sentiment inference. A comprehensive survey [13,14] indicated that when using opinion phases, topic model based methods perform better than other bag-of-words based models. Specifically, the vocabulary of a set of reviews is decomposed into two categories: head terms and modifier terms after POS Tagging processing. Each review consists of several pairs of head and modifier. For example, the phrase "nice service" is parsed into a pair of the head term "service" and the modifier term "nice". The words in modifier category can effectively infer the sentiment associated with the aspect implied by the corresponding head terms. While head terms are only responsible for aspect identification, and do not have to express any positive or negative sentiment. Moreover, it is straightforward to consider the dependence between the rating variables generating modifier terms and the topic variables producing head terms. Because reviews usually have different preferences across different aspects.

However, most existing topic models [20,21] cannot gain any benefit from the aspect ratings associated with reviews. For example, given two reviews both of which giving 3 stars overall, it is reasonable to assume on some aspects the reviewer is disappointed. But this information is generally difficult to infer these aspects from text. Even though we use bag-of-phrases and overall ratings, we still cannot tell whether modifier terms are expressing negative or positive views, because the detailed sentiment is mixed into the general overall rating. Motivated by this observation, we propose two new topic models which can simultaneously learn aspects and their ratings of reviews by utilizing aspect ratings and overall ratings. Aspect ratings are now very easy to obtain from websites like TripAdvisor[1] and Orbitz[2] website. TripAdvisor website provides the largest volume of reviews among review host websites. It holds 225 million reviews, most of which are associated with aspect ratings. None of review is without an overall rating. The problem we would like to address is predicting aspect ratings given overall ratings and text. Therefore, our model can be applied to any review data set without aspect ratings. The aspect ratings are only needed for training. Specifically, our model is based on opinion phrases which are pairs of head and modifier terms. The dependences between latent aspects and their ratings are captured by their latent variables. The aspect identification and rating inference is modeled simultaneously. We use Gibbs sampling to estimate the parameters of our models on the training data set, and maximizing a posteriori (MAP) method to predict aspect ratings on unrated reviews.

The rest of paper is organized as follows. Section 2 formulates the problem and notation we use. Section 3 proposes our model and describes the inference methods. Section 4 shows the data, the experiments and discuss experiment results. Finally we draw the conclusion in Sect. 5.

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

2 Related Work

The problem of review sentiment mining has been an attractive research topic in recent years. There are several lines of research. The early work focuses on the overall polarity detection, i.e., detecting whether a document expresses positive or negative. The author of [16] found that the standard machine learning techniques outperform human on the sentiment detection. Later, the problem of determining the reviewers sentiment with respect to a multi-point scale (ratings) was proposed in [15], where the problem was transformed into a multi-class text classification problem. Hidden Markov Model (HMM) is specially adapted to identify aspects and their polarity in Topic Sentiment Mixture model (TSM) [12]. Ranking methods are also used to produce numerical aspect scores [17].

In the literature, Latent Dirichlet Allocation (LDA) [3] based methods play a major role, because the ability of topic detection of LDA is very suitable for multi-facet sentiment analysis on reviews. MG-LDA [18,19] (Multi-Grain Latent Dirichlet Allocation) considers a review as a mixture of global topics and local topics. The global topics capture the properties of reviewed entities, while the local topics vary across documents to capture ratable aspects. Each word is generated from one of these topics. In their later work, the authors model the aspect rating as the outputs of linear regressions, and combine them into the model to aggregate relevant words in the corresponding aspect. Joint sentiment/topic model (JST) [9,10] focuses on aspect identification and its ratings prediction without any rating information available. In JST, the words of reviews are determined by the latent variables of topic and sentiment. Aspect and Sentiment Unification model (ASUM) [6] further assumes all the words in one sentence are sampled from one topic and one sentiment. CFACTS model [7] combines HMM with LDA to capture the syntactic dependencies between opinion words on sentence level. Given overall ratings, Latent Aspect Rating Analysis (LARA) [20,21] uses a probabilistic latent regression approach to model the relationship between latent aspect ratings and overall ratings. On the other hand, POS-Tagging technique is also frequently used in the detection of aspect and sentiment. The authors of [11] categorize the words in reviews into head terms and modifier terms with simple POS-Tagging methods and propose a PLSI based model to discover aspects and predict their ratings. Interdependent LDA model [13] captures the bi-direction influence between latent aspects and ratings based on the preprocessing of head terms and modifier terms. Senti-Topic model with Decomposed Prior (STDP) [8] learns different distributions for topic words and sentiment words with the help of basic POS-Tagging. Similar ideas are applied to separate aspects, sentiments, and background words from the text [23].

Our models are based on opinion phrases [11], but overcome the drawback of previous models that cannot take advantage of aspect ratings. We consider the relationship between several factors, such as overall ratings, aspect ratings, head terms and modifier terms.

3 Problem Formulation

In this section, we first introduce the problem and list notations we use in the models.

Formally, we define a data corpus of N review documents, denoted by $\mathcal{D} = \{x_1, x_2, \ldots, x_D\}$. Each review document x_d in the corpus is made of a sequence of tokens. Each review x_d is associated with an overall rating r_d, which takes an integer value from 1 to $S(S = 5)$. An aspect is a frequently commented attribute of a hotel, such as "value", "room", "location" and "service". A review consists of some text paragraphs that express the reviewers' opinions on aspects. For example, the occurrence of word "price" indicates the review comments on aspect "value". Each review is also associated with several integer aspect ratings $\{l_1, l_2, \ldots, l_K\}$, where K is the number of aspects.

Phrase: We assume each review is a set of some opinion phrases f which are pairs of head and modifier terms, i.e., $f = < h,\ m >$. In most cases, the head term h describes an aspect, and the modifier term m expresses the sentiment of the phrase. The POS-Tagging and basic NLP techniques can be used to extract phrases from raw text for each review.

Aspect: An aspect is a predefined attribute that reviewers may comment on. It also corresponds a probabilistic word distribution in topic models, which can be learned from data.

Rating: Each review contains an overall rating and may contain several aspect ratings. The rating of each review is an integer from 1 to 5. We assume that the overall ratings are available for each review, but the aspect ratings are available only in the reviews used for training. We assume that the rating is equivalent to the sentiment.

Review: A review is represented as a bag of phrases, i.e., $x_d = \{f_1, f_2, \ldots, f_M\}$.

Problem Definition: Given a collection of reviews with overall ratings and aspect ratings, the main problem is to (1) identify aspects of reviews, and (2) infer aspect ratings on the unrated reviews without aspect rating.

4 Models

In this section, we apply two generative models to identify aspects and learn their ratings by incorporating observed aspect ratings. We list the notations of the models in Table 1. We assume reviews are already decomposed into head terms and modifier terms using NLP techniques [13]. We propose two different models incorporating the aspect ratings as observed random variables.

One strong motivation is that existing topic models do not require aspect ratings of reviews during model training and consider it as an advantage. It may be

true in the past few years, since there are not many reviews containing aspect ratings. However, more and more review hosts, such as TripAdvisor and Orbitz, let reviewers to rate on predefined attribute as an option. The volume of such reviews is growing rapidly nowadays. It is reasonable to leverage the valuable information to build more precise and accurate models. To our best of knowledge, this study is the first work using aspect ratings.

Table 1. The table of notations

D	the number of reviews
K	the number of aspects
M	the number of opinion phrases
S	the number of distinct integers of ratings
U	the number of head terms
V	the number of modifier terms
z	the aspect/topic switcher
l	the aspect rating
h	the head term
m	the modifier term
r	the overall rating
θ	the topic distribution in a review
π	the aspect rating distribution for each topic
α	the parameter of the Dirichlet distribution for θ
β	the global aspect sentiment distribution
λ	the parameter of the Dirichlet distribution for β
δ	the parameter of the Dirichlet distribution for ϕ and ψ
ϕ	the head term distribution for each topic
ψ	the modifier term distribution for each sentiment

4.1 The Assumptions

We discuss some helpful assumptions for modeling. First, our models presume a flow of generating ratings and text. The reviewer gives an overall rating based on his impression and experience, then rates it on some aspects and writes some paragraphs. In the model of bag-of-phrases, the reviewer chooses a head term for an aspect on which he would like to comment, and a modifier term to express his opinion. This generation process is captured by our models.

Second, there is an interdependency between overall ratings and aspect ratings, and it varies with the numerical value of the overall rating. For example, when a user gives 5 star overall rating, it is extremely unlikely that the user

gives low ratings on any of the aspects. On the other hand, however, when a hotel receives a low overall rating, it does not necessarily get low ratings on all aspects. It is possible that the hotel still get positive feedbacks on some aspects. This usually occurs when the traveler is disappointed by a conflict, such as extra charges for unnecessary services. Inspired by this observation, we model this dependency with a multinomial distribution $P(\pi|r)$ and a global aspect sentiment distribution β conditioned on the overall rating in the following models.

Third, aspect ratings imply another interdependency, the one between aspects and sentiments [14]. Basically, it considers that different aspects have different sentiments. We explicitly introduce sentiment variables for modifier terms which are conditioned on aspect variables, so that meaningful aspects and sentiments can be learned from head and modifier terms respectively, but it avoids generating too many non-aspects.

We present two different supervised generative models. They both take aspect ratings as probabilistic variables. The aspect ratings π are merely K scores in the review on K aspects. They are observed in the training data and hence treating them as switchers is quite straightforward. An interesting observation is the distinction between the aspect rating and the phrase sentiment. They are both sentiment switchers and could be conditioned on the overall rating variable r. One is for aspects, the other is for phrases. If we assume they are both necessary and generated from the aspect sentiment distribution β and the overall rating r, then we have ARID model (Aspect and Rating Inference with the Discrimination of aspect sentiment and phrase sentiment) in Fig. 1. The interaction between π and r is through the global aspect sentiment distribution β and the overall rating r. It saves the direct dependency between them. If we assume in given the aspect k, the reviewer holds the same sentiment for all the modifier terms, the discrimination between aspect sentiment and phrase sentiment is redundant. It leads to our second model ARIM (Aspect and Rating Inference with Merging aspect sentiments and phrase sentiments).

4.2 The ARID Model

The ARID model, in Fig. 1, captures the review generation process and the interdependency between aspects and sentiments. Following conventional topic models for review analysis, we use random variables z and l to simulate the generating process of head and modifier terms respectively. The topic selection variable z is governed by a multinomial topic distribution θ. The sentiment variable l for each opinion phrase is also determined by aspect sentiment distribution β, the overall rating r, and the aspect switcher z.

Specifically, in ARID model, the variables π representing aspect ratings are shaded in the graphical representation since they are observed in the training dataset, but become latent variables for prediction over unrated reviews. The latent sentiment variable l is sampled from β_k where k is determined by the value of z. The overall rating variable r is also introduced to serve a switcher for both the aspect rating π and the phrase sentiment l. We would like to estimate

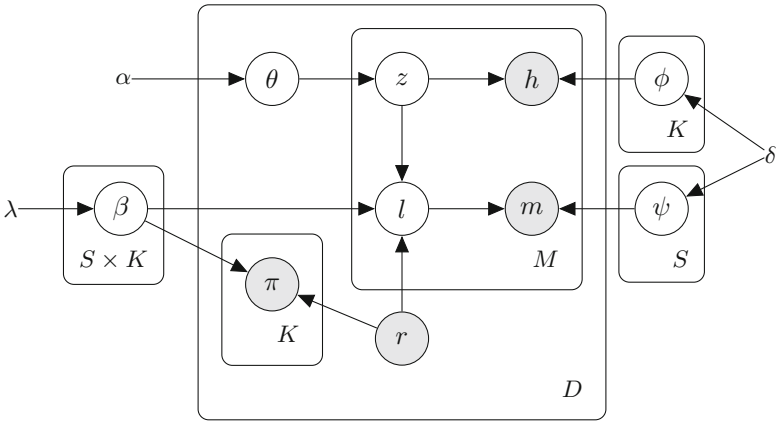

Fig. 1. Graphical Representation of ARID model. The outer box represents D reviews, while the inner box contains M phrases

the aspect rating distribution given the overall aspect sentiment distribution $p(\pi|r)$, and the latent distribution $p(l)$ and $p(z)$.

The formal generative process of our model is as follows:

- For each aspect k and each overall rating value of r
  - Sample the aspect sentiment distribution $\beta_{r,k} \sim \texttt{Dir}(\lambda)$
- For each review x_d,
  - Sample latent topic distribution variable $\theta_d \sim \texttt{Dir}(\alpha)$
  - For each aspect k from 1 to K in the review,
    - Sample aspect rating $\pi_{d,k} \sim \texttt{Mult}(\beta_{r_d,k})$
  - For each phase i from 1 to M in the review,
    - Sample aspect indicator $z_i \sim \texttt{Mult}(\theta_d)$
    - Sample sentiment indicator $l_i \sim \texttt{Mult}(\beta_{r_d,z_i})$
    - Sample head term $h_i \sim \texttt{Mult}(z_i, \phi)$
    - Sample modifier term $m_i \sim \texttt{Mult}(l_i, \psi)$

Estimation. Two parameter estimation methods are widely used for topic models, i.e., Gibbs sampling [4] and variational inference [3]. Since Gibbs sampling updating equations is relatively easy to derive and implement, for this reason, we adopt collapsed Gibbs sampling which integrates out intermediate random variables θ, ϕ, β, and ψ. For prediction, we learn the head term and the modifier term distribution ϕ, ψ, and the global aspect sentiment distribution β from z and l. The Gibbs sampling repeatedly samples latent variables $z_{a,b}$ and $l_{a,b}$ conditioned on all other latent z and l, in document a for phrase b.

The joint probability is

$$p(z,l,h,m|\alpha,\lambda,\delta,\pi,r) = \int p(\theta|\alpha)p(z|\theta)\times \\ p(h|z,\phi)p(\phi|\delta)\times \\ p(\pi|\beta,r)p(l|\beta,r,z)p(\beta|\lambda)\times \\ p(m|l,\psi)p(\psi|\delta)\ d\theta\, d\beta\, d\phi\, d\psi, \tag{1}$$

where we integrate out θ, ψ, β and ψ respectively.

We define two counters $N_{d,r,k,s,u,v}$ and $C_{d,r,k,s}$ to count the number of occurrence of opinion phrases $f_{d,i} =< h_{d,i} = u,\ m_{d,i} = v >$ and the aspect rating $\pi_{d,k}$. Specifically, $f_{d,i} =< h_{d,i} = u,\ m_{d,i} = v >$ is the phrase i of document d which has the head term u and the modifier term v. $N_{d,r,k,s,u,v}$ is the number of times that the pair of head term u and modifier term v is assigned to aspect k and sentiment s in document d, whose overall rating of the document is r. $C_{d,r,k,s}$ is the indicator of the document d that gives aspect rating s on aspect k when the overall rating of the document is r. Although given document d, its overall rating r_d is determined, we use the overall rating as a subscript for convenience.

$$N_{d,r,k,s,u,v} = \sum_{i=1}^{M} \mathbf{I}[r_d = r,\ z_{d,i} = k,\ l_{d,i} = s,\ h_{d,i} = u,\ m_{d,i} = v], \tag{2}$$

$$C_{d,r,k,s} = \mathbf{I}[r_d = r,\ \pi_{d,k} = s] \tag{3}$$

where the function $\mathbf{I}$ is the identify function. Summing out various indices results in the replacement of subscripts of N by $*$. For example,

$$N_{d,r,*,s,u,v} = \sum_{k=1}^{K} N_{d,r,k,s,u,v}. \tag{4}$$

We sample $z_{a,b}$ and $l_{a,b}$ simultaneously

$$p(z_{a,b}|z_{-(a,b)},\alpha,\delta,\lambda,h,m,r,\pi) \propto (N^{-(a,b)}_{a,r_a,z_{a,b},*,*,*} + \alpha)\times \\ \frac{N^{-(a,b)}_{*,*,z_{a,b},*,h_{a,b},*} + \delta}{N^{-(a,b)}_{*,*,z_{a,b},*,*,*} + U\delta}\times \\ \frac{N^{-(a,b)}_{*,r_a,z_{a,b},l_{a,b},*,*} + C_{*,r_a,z_{a,b},l_{a,b}} + \lambda}{N^{-(a,b)}_{*,r_a,z_{a,b},*,*,*} + C_{*,r_a,z_{a,b},*} + S\lambda}\times \\ \frac{N^{-(a,b)}_{*,*,*,l_{a,b},*,m_{a,b}} + \delta}{N^{-(a,b)}_{*,*,*,l_{a,b},*,*} + V\delta}\ . \tag{5}$$

It turns out that the aspect ratings π could be considered as pre-observed phrase sentiment counts for the global aspect sentiment distribution β. We drop

the prior parameter λ, and estimate the aspect sentiment distribution β with aspect ratings π and overall ratings r of the training data before Gibbs sampling using Eq. (6).

$$\beta_{r,k,s} = \frac{C_{*,r,k,s}}{C_{*,r,k,*}}. \tag{6}$$

The third term of the right hand of Eq. 5 is replaced by

$$\frac{N^{-(a,b)}_{*,r_d,z_{a,b},l_{a,b},*,*} + \tilde{\lambda}\beta_{r_d,z_{a,b},l_{a,b}}}{N^{-(a,b)}_{*,r_d,z_{a,b},*,*,*} + \tilde{\lambda}}, \tag{7}$$

where $\tilde{\lambda}$ is the scaling factor for β. The parameters of AIRD ψ, ϕ, θ are estimated by

$$\phi_{k,u} = \frac{N_{*,*,k,*,u,*} + \delta}{N_{*,*,k,*,*,*} + U\delta},\ \psi_{s,v} = \frac{N_{*,*,*,s,*,v} + \delta}{N_{*,*,*,s,*,*} + V\delta},\ \theta_{d,k} = \frac{N_{d,r_d,k,*,*,*} + \alpha}{N_{d,r_d,*,*,*,*} + K\alpha}. \tag{8}$$

Incorporating Prior Knowledge. We use a small set of seed words to initialize the aspect term distribution ϕ [20]. Learning the head term distribution for each aspect is difficult to converge without any prior knowledge, since each review use similar set of words for commenting on hotels. We consider the seed words as the pseudo-count which means the amount of δ words are added to $\phi_{k,u}$ by before Gibbs sampling.

Prediction. The focus of applying our model is the prediction on the unrated reviews without aspect ratings. Given an opinion phrase $f_{d,i} =< h_{d,i},\ m_{d,i} >$ and the overall rating r_d in a new document d, we identify which aspect $\hat{z}_{d,i}$ does that phrase belongs to, and predict the aspect rating $\hat{l}_{d,i}$. We drop the two subscripts d and i for simplicity. we first predict $\hat{z}$ by maximizing the posterior probability $p(z|h, m, r, \alpha, \beta, \phi, \psi)$. Using Bayes theorem, it is equivalent to maximize

$$p(z, h, m, r|\alpha, \beta, \phi, \psi) = \int p(\theta|\alpha)p(z|\theta)p(h|z, \phi)p(l|z, r, \beta)p(m|l, \psi)\, d\theta dl, \tag{9}$$

then we predict $\hat{l}$ with

$$\mathbb{E}[p(l|\hat{z}, h, m, r, \beta, \phi, \psi, \alpha)]. \tag{10}$$

The reason to consider the expectation of l is that the aspect rating is actually a numerical value, rather than a discrete category label. The importance of each possible value l is measured by its probability. The aspect weight for a new document could be learned again via Gibbs sampling, but we simply assume θ is a uniform distribution, because a review on hotel should probably comment on all the most concerned aspects. The terms in Eq. (9) we need to compute are $p(h|z, \phi) = \phi_{z,h}$, $p(l|z, r, \beta) = \beta_{r,k,l}$, and $p(m|l, \psi) = \psi_{l,m}$.

4.3 The ARIM Model

In this model, we assume the aspect sentiment is equivalent to the phrase sentiment. In other words, if all the modifier terms are categorized into the same aspect k, they share the same sentiment, i.e., the aspect sentiment. Therefore, we could just use only one sentiment indicator for both the aspect and the phrase. ARIM (Aspect and Rating Inference Merging aspect sentiments and phrase sentiments) is illustrated in Fig. 2.

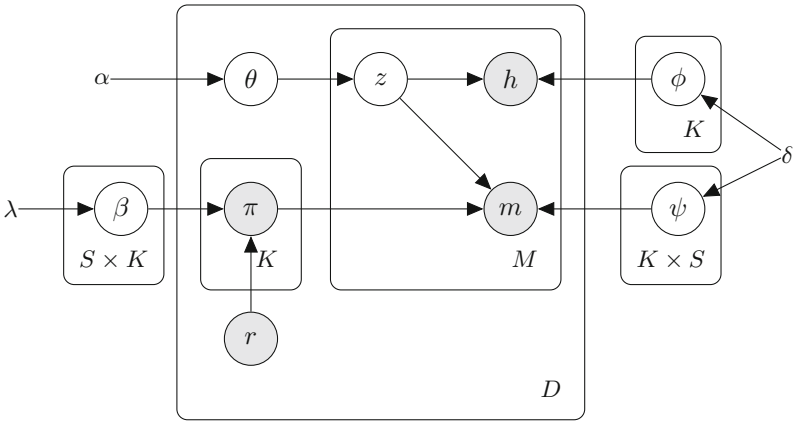

Fig. 2. Graphical Representation of the ARIM model

ARIM models aspect ratings as π like in ARID, but π is also used for phrase sentiment. The aspect ratings are available in the training data, the influence from β to m is blocked by π by d-separation theory [2] of graphical models. Therefore, the modifier term is directly determined by the aspect ratings π instead of β. In the generative procedure of ARIM, the modifier term m_i is sampled from $\psi_{z_i,\pi_{z_i}}$, and π follows a multinomial distribution with parameter β.

We still use Gibbs sampling to estimate z and β. The iterative updating function is

$$p(z_{a,b}|z_{-(a,b)},\alpha,\delta,\lambda,h,m,r,\pi) \propto (N^{-(a,b)}_{a,r_a,z_{a,b},*,*,*}+\alpha)\times \frac{N^{-(a,b)}_{*,*,z_{a,b},*,h_{a,b},*}+\delta}{N^{-(a,b)}_{*,*,z_{a,b},*,*,*}+U\delta}\times \frac{N^{-(a,b)}_{*,*,z_{a,b},\pi_{a,z_{a,b}},*,m_{a,b}}+\delta}{N^{-(a,b)}_{*,*,z_{a,b},\pi_{a,z_{a,b}},*,*}+V\delta}. \tag{11}$$

The parameters of ARIM model ϕ, θ and β is estimated by Eqs. (8) and (6). But the number of ψ is $K \times S$. It is estimated by

$$\psi_{k,s,v} = \frac{N_{*,*,k,s,*,v}+\delta}{N_{*,*,k,s,*,*}+V\delta}. \tag{12}$$

When ARIM is applied on the reviews without aspect ratings, we integrate out the latent aspect rating variable π to compute MAP $\hat{z}$ of $p(z|h, m, r, \alpha, \beta, \phi, \psi)$, which equals to

$$p(z, h, m, r|\alpha, \beta, \phi, \psi) = \int p(\theta|\alpha)p(z|\theta)p(h|z, \phi)p(m|z, \psi, \pi)p(\pi|\beta, r)\, d\pi d\theta \,. \tag{13}$$

Like Eq. (9), we again assume θ is a uniform distribution, and the terms in Eq. (13) $p(h|z, \phi) = \phi_{z,h}$, $p(m|z, r, \beta, \psi) = \sum_{s=1}^{5} \phi_{z,s,m}\beta_{r,z,s}$ by integrating out π. The estimated aspect rating $\mathbb{E}[p(\pi_k|\hat{z}, h, m, r, \beta, \phi, \psi, \alpha)]$ is computed by all the opinion phrase whose $\hat{z} = k$.

5 Experiments

In this section, we describe the review data we use and evaluate the performance of our models.

5.1 Data

The data set we use for performance evaluation is crawled from TripAdvisor [20]. Each of review in the data set is associated with an overall rating and 7 aspect ratings all within the range from 1 to 5. However some aspects such as *Cleanliness*, *Check in/front desk* are rarely rated. To better train and evaluate methods, we use only four mostly commented aspects, *Value*, *Room*, *Location* and *Service*. We only keep reviews with all four aspect ratings to evaluate and compare different models. We use NLTK [1] to tokenize the review text, remove stop words, remove infrequent words, apply POS-Tagging technique [13] to extract opinion phrases, and filter out short reviews which contains less than 10 phrases. The final data set contains 1,814 hotels and 31,013 reviews. We randomly take 80 % data as the training data set, the rest is the testing data set. The seed words used to initialize the head term distribution ϕ is in Table 2, which form a very small set of words.

5.2 Aspect Identification

In this section, we demonstrate that ARID and AIRM can identify meaningful aspects. In Table 3, we present top 3 frequentest head terms for each aspect

Table 2. Seed words

Aspect	Seed words
Value	value, fee, price, rating
Room	windows, room, bed, bath
Location	transportation, walk, traffic, shop
Service	waiter, breakfast, staff, reservation

Table 3. Frequentest head terms and modifier terms by ARIM

Aspect	Head terms	Modifier terms
Value	deal, price, charge	good, great, reasonable
Room	house, mattress, view	comfortable, clean, nice
Location	parking, street, bus	great, good, short
Service	manager, check-in, frontdesk	friendly, good, great

learned by ARIM. In other words, they have highest values in ϕ_k. We also list top 3 frequentest modifier terms for each aspect. As we can see, ARIM successfully extracted ratable aspects from reviews, and learned aspect-specific sentiment words as well. For example, "comfortable" is frequently used to describe aspect "Room", but not for other aspects. We also observe that people also like to use vague sentiment words for all aspects, such as "good", "great".

5.3 Metric

We use RMSE(Root-mean-square error)[3] to measure the performance of predicting aspect ratings for each hotel in the testing set. Letting the predicted aspect rating for hotel d on aspect k be $\hat{\pi}_{d,k}$ with ground-truth being $\pi_{d,k}$, the RMSE can be represented as Eq. (14).

$$\text{RMSE}(\hat{\pi}_{d,k}, \pi_{d,k}) = \sqrt{\frac{1}{DK}\sum_{d=1}^{D}\sum_{k=1}^{K}(\hat{\pi}_{d,k} - \pi_{d,k})^2} \tag{14}$$

RMSE measure shows how accurate one model could predicate aspect ratings. We also use Pearson correlation to describe the linear relationship between the predicted and the ground-truth aspect ratings, which is Eq. 15. π_d is the vector of the aspect ratings of document d.

$$\rho_{\text{aspect}} = \frac{1}{D}\sum_{d=1}^{D}\rho(\pi_d, \hat{\pi}_d) \tag{15}$$

Since the rating is merely an ordinal variable, whose value does not have the meaning as the numerical value. But its value has a clear ordering. Therefore, we adopt Pearson linear correlation ρ_{aspect} on the aspect ratings within each review to evaluate how a model keeps the aspect order in terms of ratings. For each aspect, it is reasonable to compute the linear correlation across hotels ρ_{hotel} as in Eq. (16). The measure is used to test whether the model could predict the order of hotels in teams of an aspect rating. π_k consists of all the aspect ratings of all the hotels on the aspect k,

$$\rho_{\text{hotel}} = \frac{1}{K}\sum_{k=1}^{K}\rho(\pi_k, \hat{\pi}_k)\ . \tag{16}$$

[3] http://en.wikipedia.org/wiki/RMSE.

5.4 Aspect Rating Prediction

In this section, we present the experiment results on the reviews without any aspect rating in Table 4. We compared three different models and one baseline. The baseline predicts all the aspect ratings of each review with the given overall rating. Since the baseline predicts the aspect ratings of a review with a constant value, $\rho_{\text{aspect}} = 0$. From the results, we observe that ARID and ARIM have close performance, but both of them outperform the baseline and LARAM [21]. The main reason is that ARID and ARIM can capture the interdependency between aspects, their ratings and modifier terms, thanks to the aspect ratings in the training data set.

Moreover, ARIM is better than ARID, which confirms our observation. The sentiment of aspect and modifier terms is not so different from each other. Reviewers hold similar attitude with different modifier terms when commenting on one aspect. Therefore, merging aspect sentiment with modifier sentiment does not deteriorate the power of the models. The information learned from the training data in ARID and ARIS is stored in β, ϕ, ψ, which are used to predict the aspect ratings in both models. ARID model has K kinds of modifier term distributions ψ; while ARIS has $K \times S$, since the modifier term m in ARIS is dependent on the aspect switcher z and the sentiment l. ARID estimates a general sentiment distribution across all aspects, but ARIM could learn aspect-specific sentiment distribution by modeling aspect-dependent sentiment. During the inference, although the aspect on which the opinion phrases comment is determined by its head term h, ARID infers the sentiment for each modifier term from a coarse sentiment distribution; while ARIM can obtain more find-grained sentiment using its $K \times S$ modifier term distributions. The ψ in ARIM fine-tunes the predicting results based on β and ϕ. Therefore, in terms of Pearson correlation metric, ARIM has better performance. In terms of ρ_{hotel}, all four approaches have similar scores. On the hotel level, the aspect ratings are averaged across all reviews, while the goals of these four methods are predicting the ratings of each individual review. The difference between each method on predicted aspect ratings for each review is small. Therefore, there is no much difference on the measure ρ_{hotel}.

Table 4. Performance of aspect inference

Measure	Baseline	LARAM	ARID	ARIM
RMSE	0.702	0.632	0.588	0.510
ρ_{aspect}	0.0	0.217	0.176	0.248
ρ_{hotel}	0.755	0.755	0.723	0.758

6 Conclusion

In this paper, we propose two models for aspect and its sentiment inference, ARID and ARIM. Both of them can employ the overall ratings and the aspect

ratings in reviews to identify the aspects on which an unrated review comments, and uncover the corresponding latent aspect ratings. The two models are based on topic models, but explicitly consider the interdependency between aspect ratings aspect terms, and sentiment terms. The opinion phrases of head terms and modifier terms are extracted by using simple POS-Tagging techniques. The most important contribution is that the two models incorporate the aspect ratings as observed variables into the models, and significantly improve the prediction performance of aspect ratings. The difference between them is whether the sentiment of modifier terms should be merged with the sentiment of aspects. Gibbs sampling and MAP is used for estimation and inference, respectively. The experiments on large hotel reviews show that ARID and ARIM have better performance in terms of RMSE and Pearson correlation. In the future, we would investigate the methods that can automatically generate ratable aspects from text, not from the predefined seed words. Another interesting research topic is to explore the relation between different aspects [5,22]. The different aspects in one review may share the similar sentiments.

Acknowledgment. The work is partially supported by National Science Foundation under grants CNS-1126619, IIS-121302, and CNS-1461926 and the U.S. Department of Homeland Security under grant Award Number 2010-ST-062-000039, the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer-Verlag New York Inc., Secaucus (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003)
4. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. U.S.A. **101**(Suppl. 1), 5228–5235 (2004)
5. Guo, Y., Xue, W.: Probabilistic multi-label classification with sparse feature learning, pp. 1373–1379, August 2013
6. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Forth International Conference on Web Search and Web Data Mining, p. 815. ACM Press, New York (2011)
7. Lakkaraju, H., Bhattacharyya, C.: Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 498–509 (2011)
8. Li, C., Zhang, J., Sun, J.T., Chen, Z.: Sentiment topic model with decomposed prior. In: SIAM International Conference on Data Mining (SDM 2013). Society for Industrial and Applied Mathematics (2013)
9. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, p. 375. ACM Press, New York, November 2009

10. Lin, C., He, Y., Everson, R., Ruger, S.M.: Weakly supervised joint sentiment-topic detection from text. IEEE Trans. Knowl. Data Eng. **24**(6), 1134–1145 (2012)
11. Lu, Y., Zhai, C., Sundaresan, N.: Rated aspect summarization of short comments. In: Proceedings of the 18th International Conference on World Wide Web, p. 131. ACM Press, New York (2009)
12. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web, pp. 171–180. ACM (2007)
13. Moghaddam, S.: ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews categories and subject descriptors. In: Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 665–674 (2011)
14. Moghaddam, S., Ester, M.: On the design of LDA models for aspect-based opinion mining. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 803–812 (2012)
15. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. pp. 115–124, June 2005
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing - EMNLP 2002, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown, July 2002
17. Snyder, B., Barzilay, R.: Multiple aspect ranking using the good grief algorithm. In: Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 300–307, April 2007
18. Titov, I., McDonald, R.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 308–316. ACL (2008)
19. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, p. 111. ACM Press, New York (2008)
20. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 783. ACM Press, New York (2010)
21. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis without aspect keyword supervision. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 618. ACM Press, New York (2011)
22. Zeng, C., Li, T., Shwartz, L., Grabarnik, G.Y.: Hierarchical multi-label classification over ticket data using contextual loss. In: 2014 IEEE Network Operations and Management Symposium (NOMS), pp. 1–8. IEEE, May 2014
23. Zhao, W., Jiang, J., Yan, H., Li, X.: Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56–65, October 2010