Quad-tuple PLSA: Incorporating Entity and Its Rating in Aspect Identification

Wenjuan Luo[1,2], Fuzhen Zhuang[1], Qing He[1], and Zhongzhi Shi[1]

[1] The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100039, China
{luowj,zhuangfz,heq,shizz}@ics.ict.ac.cn

Abstract. With the opinion explosion on Web, there are growing research interests in opinion mining. In this study we focus on an important problem in opinion mining — Aspect Identification (AI), which aims to extract aspect terms in entity reviews. Previous PLSA based AI methods exploit the 2-tuples (e.g. the co-occurrence of head and modifier), where each latent topic corresponds to an aspect. Here, we notice that each review is also accompanied by an entity and its overall rating, resulting in quad-tuples joined with the previously mentioned 2-tuples. Believing that the quad-tuples contain more co-occurrence information and thus provide more ability in differentiating topics, we propose a model of Quad-tuple PLSA, which incorporates two more items — entity and its rating, into topic modeling for more accurate aspect identification. The experiments on different numbers of hotel and restaurant reviews show the consistent and significant improvements of the proposed model compared to the 2-tuple PLSA based methods.

Keywords: Quad-tuple PLSA, Aspect Identification, Opinion Mining.

1 Introduction

With the Web 2.0 technology encouraging more and more people to participate in online comments, recent years have witnessed the opinion explosion on Web. As large scale of user comments accumulate, it challenges both the merchants and customers to analyze the opinions or make further decisions. As a result, opinion mining which aims at determining the sentiments of opinions has become a hot research topic.

Additionally, besides the simple overall evaluation and summary, both customers and merchants are becoming increasingly concerned in certain aspects of the entities. Take a set of restaurant reviews as example. Common restaurant aspects include “food”, “service”, “value” and so on. Some guests may be interested in the “food” aspect, while some may think highly of the “value” or “service” aspect. To meet these personalized demands, we need to decompose the opinions into different aspects for better understanding or comparison.

On the other hand, it also brings out perplexity for merchants to digest all the customer reviews in case that they want to know in which aspect they

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 392–404, 2012.

lack behind their competitors. As pointed out in [12], the task of aspect-based summarization consists of two subtasks: the first is Aspect Identification (AI), and the second is sentiment classification and summarization. The study in this paper mainly focuses on the first task, which aims to accurately identify the aspect terms in the reviews for certain type of entities.

Fig. 1. Sample Reviews

As shown in Figure 1, there are 3 reviews on different hotels, where the description for the same aspect is stained in the same color. One of a recent works in this area argues that it is more sensible to extract aspects from the phrase level rather than the sentence level since a single sentence may cover different aspects of an entity (as shown in Figure 1, a sentence may contain different colored terms) [5]. Thus, Lu et al. decompose reviews into phrases in the form of (*head*, *modifier*) pairs. A head term usually indicates the aspect while a modifier term reflects the sentiment towards the aspect. Take the phrase "excellent staff" for example. The head "staff" belongs to the "staff/front desk" aspect, while the modifier "excellent" shows a positive attitude to it. Utilizing the (*head*, *modifier*) pairs, they explore the latent topics embedded in it with aspect priors. In other words, they take the these 2-tuples as input, and output the latent topics as the identified aspects.

In this study, we observe that besides the *(head, modifier)* pairs each review is often tied with an entity and its overall rating. As shown in Figure 1, a hotel name and an overall rating are given for each review. Thus, we can construct the quad-tuples of

$$(head, modifier, rating, entity),$$

which indicates that a phrase of the *head* and *modifier* appears in the review for this *entity* with the *rating*. For example, the reviews in Figure 1 include the following quad-tuples,

(*price*, *good*, *5*, *Quality Inn*); (*staff*, *awesome*, *5*, *Quality Inn*);
(*location*, *good*, *4*, *L.A.Motel*); (*bed*, *small*, *1*, *Hotel Elysee*).

With these quad-tuples from the reviews for a certain type of entities, we further argue that they contain more co-occurrence information than 2-tuples, thus provide more ability in differentiating terms. For example, reviews with the same rating tend to share similar modifiers. Additionally, reviews with the same rating on the same entity often talk about the same aspects of that entity (imagine that people may always assign lowest ratings to an entity because of its low quality in certain aspect). Therefore, incorporating entity and rating into the tuples may facilitate aspect generation.

Motivated by this observation, we propose a model of Quad-tuple PLSA (QPLSA for short), which can handle two more items (compared to the previous 2-tuple PLSA [1,5]) in topic modeling. In this way we aim to achieve higher accuracy in aspect identification. The rest of this paper is organized as follows: Section 2 presents the problem definition and preliminary knowledge. Section 3 details our model Quad-tuple PLSA and the EM solution. Section 4 gives the experimental results to validate the superiority of our model. Section 5 discusses the related work and we conclude our paper in Section 6.

2 Problem Definition and Preliminary Knowledge

In this section, we first introduce the problem, and then briefly review Lu's solution–the Structured Probabilistic Latent Semantic Analysis (SPLSA) [5]. The frequently used notations are summarized in Table 1.

Table 1. Frequently used notations

Symbol	Description
t	the comment
$\mathbf{T}$	the set of comments
h	the head term
m	the modifier term
e	the entity
r	the rating of the comment
q	the quad-tuple of (h,m,r,e)
z	the latent topic or aspect
K	the number of latent topics
Λ	the parameters to be estimated
$n(h,m)$	the number of co-occurrences of head and modifier
$n(h,m,r,e)$	the number of co-occurrences of head,modifier, rating and entity
$\mathbf{X}$	the whole data set

2.1 Problem Definition

In this section, we give the problem definition and the related concepts.

Definition 1 (Phrase). *A phrase $f = (h, m)$ is in the form of a pair of head term h and modifier m. And SPLSA adopts such (head, modifier) 2-tuple phrases for aspect extraction.*

Definition 2 (Quad-tuple). *A quad-tuple $q = (h, m, r, e)$ is a vector of head term h, modifier m, rating r and entity e. Given a review on entity e with rating r, we can generate a set of quad-tuples, denoted by*

$$\{(h, m, r, e)|Phrase\,(h, m) \text{ appears with rating } r \text{ in a review of entity } e\}.$$

Aspect Cluster. An aspect cluster A_i is a cluster of head terms which share similar meaning in the given context. We represent $A_i = \{h|\mathcal{G}(h) = i\}$, where $\mathcal{G}$ is a mapping function that maps h to a cluster aspect A_i.

Aspect Identification. The goal of aspect identification is to find the mapping function $\mathcal{G}$ that correctly assigns the aspect label for given head term h.

2.2 Structured PLSA

Structured PLSA (SPLSA for short) is a 2-tuple PLSA based method for rated aspect summarization. It incorporates the structure of phrases into the PLSA model, using the co-occurrence information of head terms and their modifiers. Given the whole data $\mathbf{X}$ composed of (head, modifier) pairs, SPLSA arouses a mixture model with latent model topics z as follows,

$$p(h, m) = \sum_z p(h|z)p(z|m)p(m). \tag{1}$$

The parameters of $p(z|m)$, $p(h|z)$ and $p(m)$ can be obtained using the EM algorithm by solving the maximum log likelihood problem in the following,

$$\log p(\mathbf{X}|\Lambda) = \sum_{h,m} n(h, m) \log \sum_z p(z|m)p(h|z)p(m), \tag{2}$$

where Λ denotes all the parameters. And the prior knowledge of seed words indicating specific aspect are injected in the way as follows:

$$p(h|z; \Lambda) = \frac{\sum_m n(h, m)p(z|h, m; \Lambda^{old}) + \sigma p(h|z_0)}{\sum_{h'} \sum_m n(h', m)p(z|h', m; \Lambda^{old}) + \sigma}, \tag{3}$$

where z_0 denotes the priors corresponding to the latent topic z, and σ is the confidential parameter of the head term h belonging to aspect z_0. And each h is grouped into topic z with the largest probability of generating h, which was the aspect identification function in SPLSA: $A(h) = \arg\max_z p(h|z)$.

3 QPLSA and EM Solution

3.1 QPLSA

In SPLSA, aspects are extracted based on the co-occurrences of head and modifier, namely a set of 2-tuples. Next, we will detail our model–QPLSA, which takes the quad-tuples as input for more accurate aspect identification.

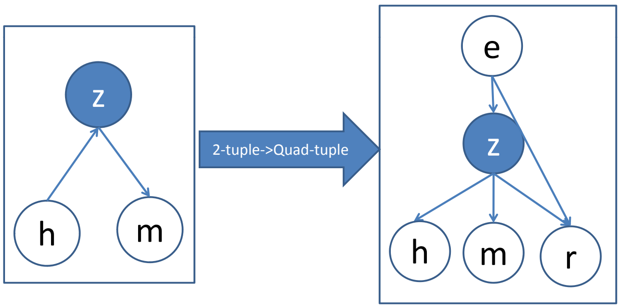

Fig. 2. From SPLSA Model to QPLSA Model

Figure 2 illustrates the graphical model of QPLSA. The directed lines among the nodes are decided by the understandings on the dependency relationships among these variables. Specifically, we assume that given a latent topic z, h and m are conditionally independent. Also, a reviewer may show different judgement toward different aspects of the same entity. Thus, rating r is jointly dependent on entity e and latent topic z. From the graphic model in Figure 2, we can write the joint probability over all variables as follows:

$$p(h, m, r, e, z) = p(m|z)p(h|z)p(r|z, e)p(z|e)p(e). \tag{4}$$

Let $\mathbf{Z}$ denote all the latent variables, and given the whole data $\mathbf{X}$, all the parameters can be approximated by maximizing the following log likelihood function,

$$\log p(\mathbf{X}|\Lambda) = \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\Lambda) \quad = \sum_{h,m,r,e} n(h, m, r, e) \log \sum_{z} p(h, m, r, e, z|\Lambda), \tag{5}$$

where Λ includes the parameters of $p(m|z)$, $p(h|z)$, $p(r|z, e)$, $p(z|e)$ and $p(e)$. The derivation of EM algorithm is detailed in next subsection.

3.2 Deriving the EM Solution

Traditionally, the Expectation-Maximization(EM) algorithm is utilized for optimization of PLSA based methods. In our model, we also adopt the EM algorithm to maximize the log likelihood function in Equation (5). Specifically, the lower bound (Jensen's inequality) $\mathcal{L}_0$ of (5) is:

$$\mathcal{L}_0 = \sum_z q(z) \log\{\frac{p(h,m,r,e,z|\Lambda)}{q(z)}\}. \tag{6}$$

where $q(z)$ could be an arbitrary function, and here we set $q(z) = p(z|h,m,r,e;\Lambda^{old})$ and substitute into (6):

$$\mathcal{L}_0 = \underbrace{\sum_z p(z|h,m,r,e;\Lambda^{old}) \log p(z,h,m,r,e|\Lambda)}_{\mathcal{L}} \\ - \underbrace{\sum_z p(z|h,m,r,e;\Lambda^{old}) \log\{p(z|h,m,r,e;\Lambda^{old})\}}_{const} = \quad \mathcal{L} + const. \tag{7}$$

E Step: Constructing $\mathcal{L}$. For the solution of (5),we have:

$$\mathcal{L} = \sum_{h,m,r,e,z} n(h,m,r,e)p(z|h,m,r,e;\Lambda^{old}) \cdot \log[p(e)p(z|e)p(h|z)p(m|z)p(r|e,z)], \tag{8}$$

where

$$p(z|e,h,m,r) = \frac{p(e)p(z|e)p(h|z)p(m|z)p(r|e,z)}{\sum_z p(e)p(z|e)p(h|z)p(m|z)p(r|e,z)}. \tag{9}$$

M Step: Maximizing $\mathcal{L}$. Here we maximize $\mathcal{L}$ with its parameters by Lagrangian Multiplier method. Expand $\mathcal{L}$ and extract the terms containing $p(h|z)$. Then, we have $\mathcal{L}_{[p(h|z)]}$ and apply the constraint $\sum_h p(h|z) = 1$ into the following equation:

$$\frac{\partial[\mathcal{L}_{[p(h|z)]} + \lambda(\sum_h p(h|z) - 1)]}{\partial p(h|z)} = 0, \tag{10}$$

we have

$$\hat{p}(h|z) \propto \sum_{m,r,e} p(z|h,m,r,e;\Lambda^{old}). \tag{11}$$

Note that $\hat{p}(h|z)$ should be normalized via

$$\hat{p}(h|z) = \frac{\sum_{m,r,e} n(h,m,r,e)p(z|h,m,r,e;\Lambda^{old})}{\sum_{h',m,r,e} n(h',m,r,e)p(z|h',m,r,e;\Lambda^{old})}. \tag{12}$$

Similarly, we have:

$$p(e) = \frac{\sum_{z,h,m,r} n(h,m,r,e)p(z|e,h,m,r;\Lambda^{old})}{\sum_{h,m,r,e} n(h,m,r,e;\Lambda^{old})}, \tag{13}$$

$$p(z|e) = \frac{\sum_{h,m,r} n(h,m,r,e)p(z|e,h,m,r;\Lambda^{old})}{\sum_{h,m,r,z'} n(h,m,r,e)p(z'|e,h,m,r;\Lambda^{old})}, \tag{14}$$

$$p(m|z) = \frac{\sum_{e,h,r} n(h,m,r,e)p(z|e,h,m,r;\Lambda^{old})}{\sum_{e,h,r,m'} n(h,m',r,e)p(z|e,h,m',r;\Lambda^{old})}, \tag{15}$$

$$p(r|z,e) = \frac{\sum_{h,m} n(h,m,r,e)p(z|e,h,m,r;\Lambda^{old})}{\sum_{h,m,r'} n(h,m,r',e)p(z|e,h,m,r';\Lambda^{old})}. \tag{16}$$

3.3 Incorporating Aspect Prior

For specific aspect identification, we may have some domain knowledge about aspects. For instance, the aspect "food" may include a few seed words such as "breakfast", "potato", "drink" and so on. Specifically, we use a unigram language model $p(h|z)$ to inject the prior knowledge for the aspect z. Take the aspect "food" as an example, we can assign the conditional probability $p(\text{breakfast}|\text{food})$, $p(\text{potato}|\text{food})$ and $p(\text{drink}|\text{food})$ with a high value of probability τ (e.g., $\tau(0 \leq \tau \leq 1)$ is a pre-defined threshold).

Similarly with the method in Lu et al. [5], we introduce a conjugate Dirichlet prior on each unigram language model, parameterized as $Dir(\sigma p(h|z) + 1)$, and σ denotes the confidence for the prior knowledge of aspect z. Specifically, the prior for all the parameters is given by:

$$p(\Lambda) \propto \prod_z \prod_h p(h|z)^{\sigma p(h|z)} \tag{17}$$

where $\sigma = 0$ if we have no prior knowledge on z. Note that adding the prior can be interpreted as increasing the counts for head term h by $\sigma + 1$ times when estimating $p(h|z)$. Therefore, we have:

$$p(h|z;\Lambda) \quad = \frac{\sum_{m,r,e} n(h,m,r,e)p(z|h,m,r,e;\Lambda^{old}) + \sigma p(h|z)}{\sum_{h',m,r,e} n(h',m,r,e)p(z|h',m,r,e;\Lambda^{old}) + \sigma}. \tag{18}$$

3.4 Aspect Identification

Our goal is to assign the head term h to a correct aspect label, and we follow the mapping function $\mathcal{G}$ as SPLSA [5]:

$$\mathcal{G}(h) = \arg\max_z p(h|z), \tag{19}$$

where we select the aspect which generates h with the largest probabilty as the aspect label for head term h.

4 Experiments

In this section, we present the experimental results to evaluate our model QPLSA. Firstly, we introduce the data sets and implementation details, and then give the experimental results in the following subsections.

4.1 Data Sets

We adopt two different datasets for evaluation, which are detailed in Table 2. The first dataset is a corpus of hotel reviews provided by Wang et al. [14]. The data set includes 246,399 reviews on 1850 hotels with each review associated with an overall rating and 7 detailed ratings about the pre-defined aspects, and the value of the rating ranges from 1 star to 5 stars. Table 2 also lists the prior knowledge of some seed words indicating specific aspects.

The other dataset is about restaurant reviews from Snyder et al. [11], which is much sparser than the previous one. This dataset contains 1609 reviews on 420 restaurants with each review associated with an overall rating and 4 aspect ratings. For both of the datasets, we decompose the reviews into phrases utilizing a set of NLP toolkits such as the POS tagging and chunking functions[1].

4.2 Implementation Details

terms and manually label them as knowledge base. Specifically, for the hotel reviews we select 408 head terms and categorize them into 7 specific aspects. While for the restaurant reviews, we select 172 head terms and label them with 4 specific aspects. The details of the categorization are summarized in Table 3, and A1 to A7 corresponds to the aspects in Table 2. Here we only evaluate the results of specific aspect identification and compare our model QPLSA with SPLSA.

Table 2. Pre-defined Aspects and Prior Knowledge

Hotel Reviews		
Aspects	Prior Words	Aspect No.
Value	value,price,quality,worth	A1
Room	room,suite,view,bed	A2
Location	location,traffic,minute,restaurant	A3
Cleanliness	clean,dirty,maintain,smell	A4
Front Desk/Staff	staff,check,help,reservation	A5
Service	service,food,breakfast,buffet	A6
Business	business,center,computer,internet	A7
Restaurant Reviews		
Food	food,breakfast,potato,drink	A1
Ambience	ambience,atmosphere,room,seat	A2
Service	service,menu,staff,help	A3
Value	value,price,quality,money	A4

[1] http://opennlp.sourceforge.net/

Table 3. Aspect Identification Accuracy on Two Datasets

	Hotel Reviews								Restaurant Reviews					All
	A1	A2	A3	A4	A5	A6	A7	A1-7	A1	A2	A3	A4	A1-4	All
Categorized	52	108	93	35	39	64	17	408	73	32	42	25	172	580
QPLSA	29	69	45	21	31	47	12	**254**	29	21	23	22	**95**	**349**
SPLSA	29	61	46	20	28	46	4	234	4	0	7	5	16	250
Q-accuracy	0.56	0.64	0.48	0.60	0.79	0.73	0.71	**0.62**	0.39	0.66	0.55	0.88	**0.55**	**0.60**
S-accuracy	0.56	0.56	0.49	0.57	0.72	0.72	0.24	0.57	0.05	0	0.17	0.2	0.09	0.43

4.3 Experimental Results

Aspect Identification. We present the accuracy of aspect identification of all the head terms in Table 3. Since we focus on specific aspect extraction, our discussions only detail the results on specific aspects. In the table, Ai denote the i-th specific aspect as described in Table 2, and "A1-7" and "A1-4" denote the sum of the specific aspects for hotel reviews and restaurant reviews, respectively.

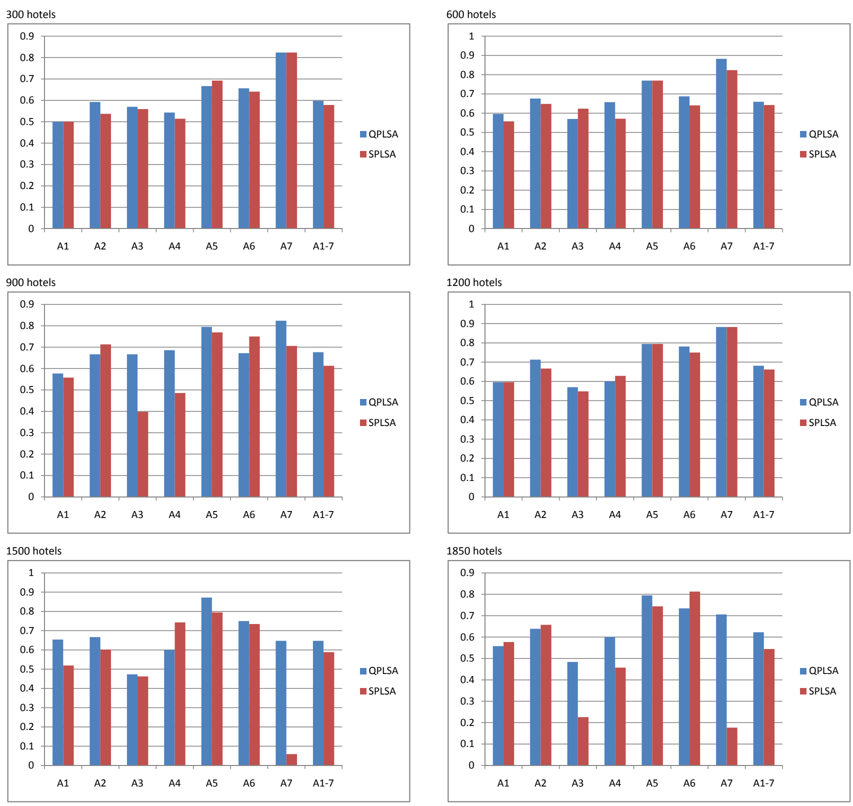

Fig. 3. Accuracy on different numbers of hotels

In Table 3, Q-accuracy denotes the accuracy of QPLSA, and S-accuracy represents that of SPLSA. From the results reported in Table 3, apparently, QPLSA achieves better performance compared to SPLSA. As can be seen, the accuracy of QPLSA for all the reviews is much higher than that of SPLSA, which indicates that quad-tuples exploits more information for specific aspect generation as opposed to 2-tuples. All the experimental results demonstrate the effectiveness of incorporating entity and its rating for aspect identification.

To further validate the superiority of QPLSA over SPLSA, we conduct systematic experiments on different data sets of hotel reviews for comparison. We carry out experiments on different numbers of hotels (e.g., 300, 600, 900, 1200, 1500 and 1850), and all the results are shown in Figure 3.

As illustrated in Fig. 3, in particular, the performance of QPLSA varies for different aspects due to the skrewness of corpse over specific topics. Nevertheless, for different numbers of hotels, that the overall accuracy of QPLSA always outperforms that of SPLSA strongly supports that Aspect Identification of QPLSA can benefit from the additional information of entity and its rating.

Representative Term Extraction. Table 4 lists representative terms for the 7 specific aspects of hotel reviews and the 4 aspects of the restaurant reviews. For each aspect, we choose 20 head terms with the largest probability, and the terms that are correctly associated with the aspects are marked with bold and italic.

Table 4. Representative terms for Different Aspects

Hotel Reviews		
Aspects	Representative Terms By QPLSA	Representative Terms By SPLSA
Value	hotel location experience ***value price*** size vacation ***rates choice deal*** job way surprise atmosphere ***quality selections money*** holiday variety spots	walk ***value price rates*** side york parking station tv orleans ***quality*** distance standards screen light ***money*** end ***charge*** line bus
Room	***room bed view pool bathroom suits*** ocean ***shower style space feel window facilities*** touch ***balcony chair bath amenities pillows furnished***	***room*** quarters area ***bed view pool*** transportation ***bathroom suits*** towels ***shower*** variety lobby ***space window facilities balcony chair bath*** sand
Location	***places restaurants area walk resort beach city street shopping minutes bus distance quarters building tourist store tour*** lobby attractions cafe	time ***restaurants*** day night ***resort trips beach*** doors ***street way minutes*** years week hour ***visit*** weekend ***block island*** evening morning
Cleanliness	***water decor towels*** fruit ***tub air*** appointed sand ***cleaning smell maintained noise music*** club ***condition*** garden republic done design francisco	floor level ***water*** flight ***air noise music*** class worlds ***cleaning smell maintained condition*** wall francisco car eggs anniversary ***notch*** afternoon
Front Desk	***staff reservation guests checking manager*** house ***airporter receptions desk help*** island eggs lady ***attitude smiles*** lounge museum kong man ***concierge***	***staff desk*** people ***guests checking*** person couples ***manager*** fun lounge children ***member receptions*** towers guys ***reservation*** cart trouble ***attitude*** lady
Service	***service breakfast food bar drinks buffet*** tv ***coffee meals wine bottle items dinner juice tea snacks dish*** screen car shuttle	***service breakfast food*** access ***bar*** tub shuttle ***drinks buffet coffee meals fruit wine bottle*** connected weather ***juice beer tea snacks***
Business Service	floor ***access internet*** side parking station standards light end class ***line sites*** wall stop ***business connected center*** district towers level	shopping problem building complaints ones ***internet*** traveller points bit tourist store cafe deal thing attractions issue star ***sites*** items city
Total	**89 correct terms**	64 correct terms
Restaurant Reviews		
Food	***food potato sauce ribs wine taste drinks fries*** parking fee dogs ***toast breakfast bun cajun pancakes croissants lasagna pies cinnamon***	***food potato sauce ribs wine sause taste drinks*** gravy diversity reduction ***feast charcoal*** plus brats nature ***tiramisu cauliflower goods***
Ambience	***atmosphere style*** cheese shrimp ***room seated music*** tomatoes ***decor game dressing*** tip orders onion mushroom garlic cocktail ***setting piano*** mousse	***atmosphere*** area ***style room seated feeling music manner piano band poster arts cello movie blues appearance*** folk medium francisco avenue
Service	***service staff menu wait guy guests carte*** chili ***attitude*** space downtown section become women ***employees critic*** poster market ***waitstaff office***	***help service staff menu attitude guests*** gras mousse maple behavior tone lettuce defines future excuse smorgasbord sports ***networkers*** supper grandmothers
Value	***priced value quality*** done management legs anniversary ***rate money*** thought cafeteria informed croutons bags elaine system bomb ***proportions recipes buy***	***priced value quality*** parking ***rate money*** ravioli ***fee*** pupils flaw heron inside winter education aiken standbys drenched ***paying*** year-old-home veteran
Total	**47 correct terms**	42 correct terms
All	**136 correct terms**	108 correct terms

Totally, for the 7 aspects of hotel reviews, there are 105 head terms accurately selected by QPLSA compared to 64 by SPLSA. Also for the 4 aspects of restaurant reviews, more correct words are captured by QPLSA than SPLSA. In all, QPLSA extracts 136 correct terms compared to 108 of SPLSA. All these results demonstrate that incorporating entity and its rating for aspect identification(or extraction) is effective.

Note that both QPLSA and SPLSA obtain much better results on dataset hotel reviews than those on restaurant reviews. The reason is that both methods are based on generative model that models the co-occurrence information. As we know, hotel review dataset is much more dense, and thus can provide enough co-occurrence information for learning.

5 Related Work

This section details some interesting study that is relevant to our research. Pang et al. [8] give a full overview of opinion mining and sentiment analysis, after describing the requests and challenges, they outlined a series of approaches and applications for this research domain. It is pointed out that sentiment classification could be broadly referred as binary categorization, multi-class categorization, regression or ranking problems on an opinionated document.

Hu and Liu [2] adopt association mining based techniques to find frequent features and identify the polarity of opinions based on adjective words. However, their method did not perform aspect clustering for deeper understanding of opinions. Similar work carried out by Popescu and Etzioni [10] achieved better performance on feature extraction and sentiment polarity identification, however, there is still no consideration of aspects.

Kim et al. [3] developed a system for sentiment classification through combining sentiments at word and sentence levels, however their system did not help users digest opinions from the aspect perspective. More approaches for sentiment analysis could be referred to [9,13,15,7], although none of these methods attach importance to aspects.

Topic models [14,4,6,5] are also utilized to extract aspects from online reviews. Lu et al. adopt the unstructured and structured PLSA for aspect identification [5], however, in their model, there is no consideration of rating or entity in the aspect generation phase. Wang et al. [14] proposed a rating regression approach for latent aspect rating analysis on reviews, still in their model they do not take account of entity. Mei et al. [6] defined the problem of topic-sentiment analysis on Weblogs and proposed Topic-Sentiment Mixture(TSM) model to capture sentiments and extract topic life cycles. However, as mentioned before, none of these topic models extracts aspects in view of quads.

A closely related work to our study could be referred to Titov and McDonald's [12] work on aspect generation. They construct a joint statistical model of text and sentiment ratings, called the Multi-Aspect Sentiment model(MAS) to generate topics from the sentence level. They build local and global topics based on the Multi-Grain Latent Dirichlet Allocation model (MG-LDA) for better aspect generation. One recent work [4] by Lakkaraju et al. also focused on sentence level aspect identification. However, according to our observation, a single sentence may address several different aspects and therefore we generate aspects from the phrase level, while they extract topics from the sentence level. Moreover, in their model, there is no consideration of entity.

6 Conclusion

In this paper, we focus on aspect identification in opinion mining and propose a quad-tuple PLSA based model which novelly incorporates the rating and entity for a better aspect generation. Compared to traditional 2-tuple(head, modifier) PLSA based modeling methods, our model exploits the co-occurrance information among quad-tuples(head, modifier, rating, entity) and extract aspects from a finer grain. After formally describing our quad-tuple PLSA(QPLSA) and applying the EM algorithm for optimization, we carry out systematic experiments to testify the effectiveness of our algorithm. Experimental results show that this method achieves better performance in aspect identification and representative term extraction compared to SPLSA(a 2-tuple PLSA based method). Our future work will focus on aspect rating prediction and sentiment summarization.

Acknowledgement. We would like to thank Ping Luo from the H.P.(Hewlett-Packard) Labs China. Besides, our work is supported by the National Natural Science Foundation of China (No. 60933004, 60975039, 61035003, 60903141, 61072085), National Basic Research Priorities Programme (No.2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06).

References

1. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd International Conference on Reserach and Development in Inforamtion Retrieval, SIGIR 1999 (1999)
2. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2004), pp. 168–177 (2004)
3. Kim, S.M., Hovy, E.: Determining the sentiment of opinors. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 1367 (2004)
4. Lakkaraju, H., Bhattacharyya, C., Bhattacharya, I., Merugu, S.: Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In: Proceedings of 2011 SIAM International Conference on Data Mining (SDM 2011), pp. 498–509 (April 2011)
5. Lu, Y., Zhai, C., Sundaresan, N.: Rated aspect summarization of short comments. In: Proceedings of the 18th International Conference on World Wide Web (WWW 2009), pp. 131–140 (2009)
6. 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 World Wide Web Conference (WWW 2007), pp. 171–180 (2007)

7. Morinaga, S., Tateishi, K.Y.K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 341–349 (2002)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. In: Foundatoins and Trends in Information Retrieval, Rome, Italy, pp. 1–135 (September 2008)
9. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86 (2002)
10. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 339–346 (2005)
11. Snyder, B., Barzilay, R.: Multiple aspect ranking using the good grief algorithm. In: Proceedings of the Joint Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies, pp. 300–307 (2007)
12. Titov, I., McDonald, R.: A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Meeting of Association for Computational Linguistics (ACL 2008), pp. 783–792. Morgan Kaufmann, Rome (2008)
13. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Meeting of Association for Computational Linguistics (ACL 2002), pp. 417–424 (2002)
14. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis on review text data: A rating regression approach. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD 2010), pp. 783–792 (2010)
15. Zhuang, L., Jing, F., Zhu, X.Y.: Movie review mining and summarization. In: Proceedings of the 15th Conference on Information and Knowledge Management (CIKM 2006), pp. 43–50 (2006)