Send Orders for Reprints to reprints@benthamscience.ae

Aspect Clustering Combined N-gram for Reviews

Shibo Zhang* and Xiaojie Wang

School of Computer Science and Technology, Beijing University of Posts and Telecommunications, Haidian District, Beijing, 100876, China

Abstract: With the increase in popularity of e-commerce, more and more customer reviews are available online, it's usually hard to go through each of them. Latent Dirichlet Allocation (LDA) was used to mine product aspect. Considering the weakness of standard LDA when processing review text, we defined the product aspect model, and predefined aspects for different domains, and proposed aspect model combined N-gram based on sentence, which can automate aspect clustering. Our experimental results show that the proposed model can cluster mostly aspects and recognize representative words for the aspect with more than one word, and achieve better sentence-level aspect precision than previously proposed aspect models.

Keywords: Aspect model, N-gram, reviews.

1. INTRODUCTION

Nowadays, there are large amount of reviews for products available online, ranging from books, restaurants, electronic devices to many others. In the review, customer will give some opinions on different aspects. Different customers may talk about on different aspects for the same product. A user who is looking for computer may care about its computational power and maximum throughput, while other cares more about graphical capability.

As we can see, there is one big problem for review mining, aspect extraction. The main problem in the context of aspect extraction is to identify those text passages which refer to mentions of product aspects. Given a dictionary of relevant product aspects, the task would be relatively easy. However, if the relevant product aspects are not known a priori, we need to derive them by examining the provided collection of review documents. We thus need to devise methods that automatically extract a set of the most relevant product aspects from a corpus of reviews.

In this paper, we proposed a aspect model that can jointly model aspects and N-Gram. We learn the aspect for each sentence level. This assumption is reasonable for review mining, since each sentence is likely related to only one aspect.

2. PREVIOUS WORK

There have been quite a lot of efforts put into aspect mining. Turney [1] uses an unsupervised algorithm based on mutual information to classify the semantic in the word and phrase level, and Choi *et al.* [2] uses conditional random fields to classify the sentiments.

The earliest attempts for detecting aspects are based on frequently occurring noun phrases [3]. This approach works well when aspect are strongly tied to the single word, but less useful when aspects uses many low frequency terms. One common solution is to use clustering techniques to group the terms that associated with the same aspects. After that, they search for opinions associated with those aspects.

With the popularity of topic model (Latent Dirichlet Allocation, LDA [4]), there are many variations of LDA model have been applied to the product review mining. Titov *et al.* [5] propose a model to model two types of topics in reviews which are global topic and local topic. The global topic is related to a global property of review, the local topic is related to the product aspects. This is because aspects are fundamentally different from the global property of products. Takuya proposed a two-level learning approach to estimate aspects [6]. Brody and Elhadad [7] proposed a local LDA model, where they extract the aspects from sentence instead of the whole review text. Wang *et al.* [8] uses bootstrap methods to extract the aspects first, and then use latent models to analyze the opinions.

All of their approaches are based on bag of words assumption that use unigram models. While unigram model is simple and computationally efficient, it lack of capability to find out meaningful phrase which consists of more than one word.

3. N-GRAM TOPIC MODEL

N-gram topic model was proposed by Wallach [9]. While the bi-gram model always generate bi-grams, the N-gram model is flexible enough that can generate phrase that consists of arbitrary length of words. The main idea is to introduce new latent variable X, which is the indicator variable to denotes whether we need to combine the current word with previous one to make bi-gram.

The whole generative process looks like this:

1874-110X/14 2014 Bentham Open

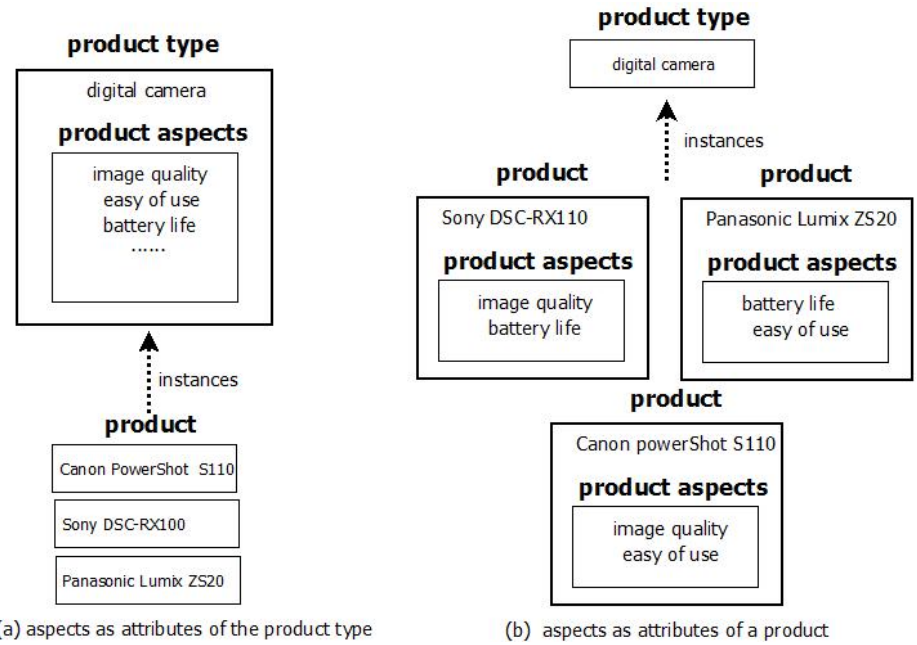

Fig. (1). The relation between the concepts product type, product and product aspect. Product aspects can be modeled as attributes of the product type (**a**) or of a concrete product (**b**).

1) *Draw Discrete distribution φ_k from Dirichlet prior β, for each topic k.*
2) *Draw Bernoulli distribution ψ_{kw} from a Beta prior γ for each topic k and each word w.*
3) *Draw Discrete distribution σ_{kw} from Dirichlet prior η for each topic k and each word w.*
4) *For each document d, for d = 1,....D*
   a) *draw document distribution θ_d from Dir(α)*
   b) *for each word i in document d*
      i. *Draw z_i from multinomial(d)*
      ii. *Draw indicator variable x_i from Bernoulli$\psi_{z_{i-1},w_{i-1}}$*
      iii. *Draw word w_i from φ_{z_i} if x = 0, else draw w_i from$\sigma_{z_i,w_{i-1}}$*

As we notice, we draw each word either from unigram model or bi-gram model, depends on the indicator variable. This is very intuitive and makes model to generate any meaningful phrases.

4. PROPOSED SENTENCE N-GRAM MODEL

4.1. Model Product Types and Aspects

As the section title implies, we distinguish between the terms product type, product, and product aspect. We refer to a product type when addressing a whole class of products. For example, the classes of digital cameras, mp3 players, hotels, or restaurants all constitute a separate product type. We use the term product to denote individual instances of a product type, *e.g.*, "Canon EOS 600D", "SanDisk Sansa Clip+". As illustrated in Fig. (**1**), product aspects may be defined with respect to a concrete product (*e.g.*, "Canon EOS 600D") or in terms of a whole product type (*e.g.*, digital cameras). How to model the aspects depends on the actual application scenario and the requirements for a review mining system. Due to the shape of our review corpora, we decided to model product aspects with regard to the product type.

4.2. Sentence N-gram Model

Standard Latent Dirichlet Allocation does not work well for short text [4], like reviews, comments and etc, is inappropriate to detect aspects in reviews, in that it tends to get global topic while rateable aspects related to reviews is more important [5]. If we apply the classical unigram or N-gram model to short text like reviews, it tends to give us higher level aspects. The problem was mentioned by Samuel Brody [7]. Now, we proposed a N-gram model based on sentence, where we assume that all the words within one sentence comes from the same aspect. We split the text into sentences based on conjunction symbols. In this model, we set N to be 2.

Fig. (**2**) shows the sentence N-gram topic model.

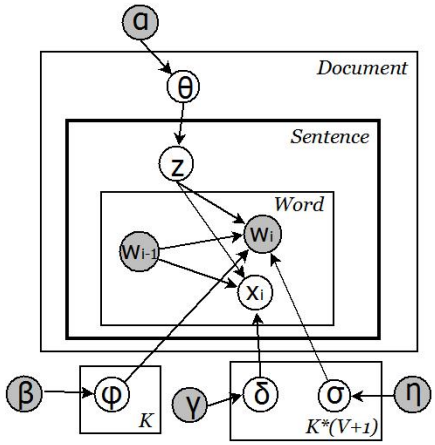

Fig. (2). Sentence based N-gram topic model.

The generative process is as follows:

1) *Draw Discrete distribution φ_k from Dirichlet prior β, for each topic k.*
2) *Draw Bernoulli distribution ψ_{kw} from a Beta prior γ for each topic k and each word w.*
3) *Draw Discrete distribution σ_{kw} from Dirichlet prior η for each topic k and each word w.*
4) *For each document d, for d = 1,....D*
   a) *draw document distribution θd from Dir(α)*
   b) *for each sentence m in document d*
      I. *Draw zm from multinomial(θd)*
      II. *For each word w_i in sentence m*
         i. *Draw indicator variable x_i from Bernoulli $\psi_{z_m,w_{i-1}}$*
         ii. *Draw word w_i from φzm if x = 0, else draw w_i from$\sigma z_m,w_{i-1}$*

The estimation of latent variables is troublesome, Gibbs sampling [10] is used to derive them, the MCMC approach plays an important role. The latent variables can be estimated as follows:

We first set that

$$c(\alpha)=\frac{\sum_i \Gamma(\alpha_i)}{\Gamma(\sum_i \alpha_i)} \tag{1}$$

For collapsed gibbs sampling, we just need to calculate the following quantity:

$$p(x_i,z_i \mid \alpha,\beta,\delta,x,w,\gamma,z_{-i}) = \frac{p(x,z,w \mid \gamma,\delta,\alpha,\beta)}{p(x_{-i},z_{-i},w \mid \gamma,\delta,\alpha,\beta)} \tag{2}$$

For the diving term, we can expand it to

$$p(x,z,w \mid \gamma,\delta,\alpha,\beta) = p(z \mid \alpha)p(x \mid w,\gamma,z)p(w \mid x,\beta,z,\delta) \tag{3}$$

The first term $p(z \mid \alpha)$ remains the same as in LDA, so we have

$$p(z \mid \alpha)=\prod_{d=1}^{D}\frac{c(\alpha+\sum_{j:\delta_j=d} I_K(z_j))}{c(\alpha)} \tag{4}$$

for $p(w \mid z,\beta)$, we need to consider two different cases.

When x_i=0

$$\begin{aligned} p(w \mid \beta,z) &= \int p(\phi \mid \beta)p(w \mid \phi,z)d\phi \\ &= \int \prod_{k=1}^{K}\prod_{i:z_i=k,x_i=0}\prod_{v=1}^{V}\phi_{k,v}^{I_{w_i=v}} \times \prod_{k=1}^{K}\frac{1}{c(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1}d\phi \\ &= \int \prod_{k=1}^{K}\frac{1}{c(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\sum_{i:z_i=k,x_i=0} I_{w_i=v}+\beta_v-1}d\phi \\ &= \prod_{k=1}^{K}\frac{c(\beta+\sum_{i:z_i=k,x_i=0} I_V(w_i))}{c(\beta)} \end{aligned} \tag{5}$$

when x_i=1

$$\begin{aligned} p(w \mid \beta,z) &= \int p(\sigma \mid \beta)p(w \mid \sigma,z)d\sigma \\ &= \int \prod_{k=1}^{K}\prod_{w=0}^{V}\prod_{i:z_i=k,w_{i-1}=w,x_i=1}\prod_{v=1}^{V}\phi_{k,v}^{I_{w_i=v}} \times \prod_{k=1}^{K}\prod_{w=0}^{V}\frac{1}{c(\beta)}\prod_{v=1}^{V}\sigma_{k,w,v}^{\delta_v-1}d\sigma \\ &= \prod_{k=1}^{K}\prod_{w=0}^{V}\frac{c(\delta+\sum_{i:z_i=k,w_{i-1}=w,x_i=1} I_V(w_i))}{c(\delta)} \end{aligned} \tag{6}$$

For the last term, we have

$$\begin{aligned} p(x \mid z,w,\gamma) &= \int p(x \mid z,w,\psi)p(\psi \mid \gamma)d\psi \\ &= \int \prod_{k=1}^{K}\prod_{w=0}^{V}\prod_{i:z_i=k,w_{i-1}=w}\prod_{s=0}^{1}\phi_{k,w,s}^{I_{x_i=s}} \times \prod_{k=1}^{K}\prod_{w=0}^{V}\prod_{s=0}^{1}\psi_{k,w,s}^{\gamma_s-1}d\psi \\ &= \prod_{k=1}^{K}\prod_{w=0}^{V}\frac{c(\gamma+\sum_{i:z_{i-1}=k,w_{i-1}=w,x_i=1} I_s(x_i))}{c(\gamma)} \end{aligned} \tag{7}$$

Finally we can get

$$\begin{aligned} p(z,x,w \mid \alpha,\beta,\gamma,\delta) &= \prod_{d=1}^{D}\frac{c(\alpha+\sum_{j:\delta_j=d} I_K(z_j))}{c(\alpha)} \times \\ &\prod_{k=1}^{K}\frac{c(\beta+\sum_{i:z_i=k,x_i=0} I_V(w_i))}{c(\beta)} \times \\ &\prod_{k=1}^{K}\prod_{w=0}^{V}\frac{c(\delta+\sum_{i:z_i=k,w_{i-1}=w,x_i=1} I_V(w_i))}{c(\delta)} \times \\ &\prod_{k=1}^{K}\prod_{w=0}^{V}\frac{c(\gamma+\sum_{i:z_i=k,w_{i-1}=w} I_S(x_i))}{c(\gamma)} \end{aligned} \tag{8}$$

5. EXPERIMENTS

Given a review document, the core task of an aspect-oriented customer review mining system is to extract all mentions of product aspects the reviewer has commented on. We designed experiment to verify the validity of our Sentence n-gram model.

5.1. Experiment Data

All the experiments we present in the following chapters are conducted on two different datasets we sampled from a web crawl of prominent customer review sites. We crawled a document collection of 417,170 hotel reviews from the travel website Tripadvisor.com and 180,911 digital camera reviews from the online retailer Amazon.com.

Our primary motivation to conduct experiments on two distinct datasets is to increase the significance and reliability of our evaluation results. In particular, being able to compare results obtained from two different domains allows us to analyze the generalizability of derived assertions more soundly. It prevents us from drawing conclusions based on phenomena that might be inherent only to one specific domain.

We select the specific domains of hotel and digital camera reviews with the following intentions in mind: First, they represent two quite distinct genres of customer reviews, namely reviews of products (*e.g.*, cars, mp3 players, refrigerators) and reviews of services (*e.g.*, restaurants, hairdressers, health clubs). Among both genres, digital cameras and hotels are very popular targets of customer reviews. Second, both domains have been considered in other, related studies, which makes our results more comparable. And third, as a very practical consideration, due to the relative popularity of the selected domains, it is easier to crawl huge collections of review documents.

5.2. Preprocessing

First, stop words are removed, Second, documents are split into sentences by ".", "?" and "!". From the acquired web crawls we randomly sampled a set of 323 hotel reviews as well as a set of 396 digital camera reviews.

Table **1** shows the detailed statistics for each dataset. Both datasets are composed of roughly 3,500 sentences and 60,000 tokens.

Table 1. Descriptive statistics of the review corpora.

Statistic	Hotel	Digital Camera
Reviews	323	396
Sentences	3512	3481
Avg.sentence/review	10.87	8.79
Min.sentence/review	1.00	1.00
Max.sentence/review	58.00	89.00
Avg.tokens/sentence	18.51	17.86
Min.tokens/sentence	1.00	1.00
Max.tokens/sentence	87.00	83.00

We predefined topics (pre-topic) for corpora, in that case, the annotator fills the "topic attribute" by selecting one of the predefined topics compiled for the hotel review and digital camera review domains. If the sentence covers multiple different topics, all associated topics are enumerated by means of a comma-separated list. In case the sentence is no-topic, that is, it cannot be associated with one of the predefined topics, the topic attribute remains empty. For the domain of hotel reviews we distinguish 10 topics, the number of topics in the camera domain is higher with 15 topics that are presented in Table **2**.

5.3. Evaluation

During formula derivation in LDA, the α and β are always set by experience points, in our experiments we set them to be 50/K and 0.01 respectively [11].

Results are shown in Table **3**. We see that within the hotel review corpus 2,636 of 3,512 sentences (75.06%) are pre-topic. In the digital camera corpus we count 2,511 of 3,481 sentences (72.13%) which are pre-topic. These numbers confirm our assumption that customer reviews are generally very focused documents and most of the provided information is relevant to the discussed product or one of its aspects.

Table 2. List of predefined product aspects for the domains of hotel and digital camera reviews.

	Hotel Domain	Digital Camera Domain
Aspect	price	connectivity
	room	accessory
	sleep quality	battery
	service	features
	location	memory
	internet	ease of use
	facility	flash
	dining	optics
	decoration	appearance
	cleanliness	picture quality
		price
		screen
		video recording
		speed

Table **3** lists the distribution of aspects for both corpora. In the hotel review domain, reviewers most often comment on the topics "price", "room", and "location". These three topics alone account for roughly 50% of all sentences in the corpus. Within the digital camera domain, the top five topics "picture quality", "memory", "price", "screen" account for around 50% of all sentences. We found our model could identify the mainly aspects on both corpus.

Table **4** shows the representative words for some aspects, these words had the commonly meaning accounting for these domains.

In hotel reviews, we predefined the aspect "sleep quality", the same "video recording" in camera reviews. Both them has a phase more than one words, it is not recognized with unigram model, because of the out-of-order words in bag of words.

Fig. (**3**) shows the results for evaluation of both datasets with three aspect model(LDA, Local LDA and Sentence N-gram). Compared to the stantard LDA baseline, the precision of our model is around 8 percentage points higher with regard to the hotel corpus (Sentence N-gram: 0.754, Standard LDA: 0.671) and approximately 9 percentage points when considering the camera corpus (Sentence N-gram: 0.737, Standard LDA: 0.643). The figure shows that both the sentence step and N-Gram step have a positive effect on the precision.

The predictive power is a important evaluation criteria when determining a model's quality. A prediction task should be defined, the performance of the model prediction

Table 3. The distribution of topics in the review corpora.

Hotel Domain		Digital Camera Domain	
Topic	**Frequency**	**Topic**	**Frequency**
price	712(20.27%)	picture quality	534(15.34%)
room	603(17.17%)	memory	479(13.76%)
location	429(12.22%)	price	462(13.27%)
dining	403(11.48%)	screen	315(9.05%)
service	331(9.42%)	speed	309(8.88%)
facility	228(6.49%)	ease of use	247(7.1%)
internet	109(3.1%)	video recording	231(6.64%)
cleanliness	107(3.04%)	optics	200(5.75%)
sleep quality	89(2.53%)	appearance	170(4.88%)
decoration	45(1.28%)	battery	90(2.59%)
		accessory	78(2.24%)
		connectivity	60(1.72%)
		flash	51(1.47%)
		features	30(0.86%)
Pre-topic	2636(75.06%)	Pre-topic	2511(72.13%)
Off-topic	876(24.94%)	Off-topic	970(27.87%)

Table 4. List of representative words for some aspect about hotel and digital camera reviews.

Aspect	Representative Words	Aspect	Representative Words
room	Window, huge, door, quiet, bed, floor, size, tv, space, phone	Price	Worth, money, save, well, price, spend, notworth, cost, well
location	Park, city, airport, distance, area, location, bus, center, walk, noise	Battery	battery, life, hours, time, cell, last, fast, hot, little
service	Service, friendly, morning, help, nothing, chair, people, time	Screen	Bright, display, color, clear, lcd, reflect, light, view, sharp
price	Good, cheap, expensive, 5star, buy, taxi, little, happy, worthy	video recording	Memory, sharp, video, space, hour, huge, shake, screen, worth
sleep quality	Bed, floor, lamp, curtain, thick, house, grass, morning		

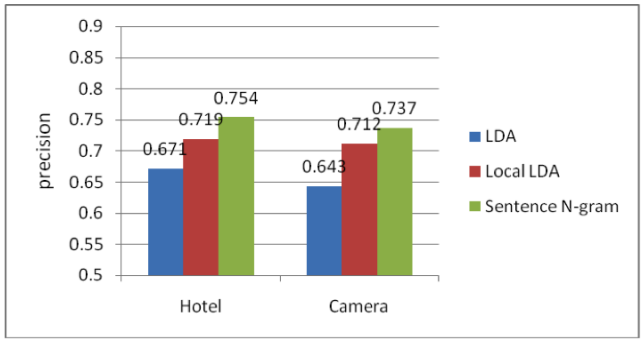

Fig. (3). Evaluation of three aspect model.

would be dependent on the task, then we can run the experiment and get the predictive power of our model. Observe that in comparison to other studies, which for example only consider four or five distinct topics [12], the coverage of our topic sets is much higher.

CONCLUSION

We presented Sentence-LDA, a model which combines LDA with N-Gram to extract aspects. In our experiments with terminology extraction, we experimented with two data sets to group product aspect. We define the relevance of identified product aspects with regard to a whole class of

products (*e.g.*, hotels or digital cameras), instead of a single specific product or model. This becomes manifest in our use of a separate, very large foreground corpus, as well as in the composition of our evaluation corpora. In contrast, the major share of related work, such as Ferreira *et al.* [13] or Hu and Liu [3], define relevance towards an individual product.

In future work, more techniques will be developed to extract aspect, and improve the accuracy of grouping product aspects, and we also will apply the proposed Sentence N-Gram model approaches for other domains.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 90920006)

REFERENCES

[1] Turney P, Littman M L, "*Unsupervised learning of semantic orientation from a hundred-billion-word corpus*," Technical Report NRC Technical Report ERB-1094, Institute for Information Technology, National Research Council Canada, (2002). 2002.

[2] Choi Y, Cardie C,and Riloff E, "Identifying sources of opinions with conditional random fields and extraction patterns," *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005,pp.355-362.

[3] Hu M, Liu B, "Mining and summarizing customer reviews," *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp.168-177.

[4] Blei D M, Ng A Y, and Jordan M I. "Latent dirichlet allocation," *the Journal of machine Learning research*, 2003, 3,pp. 993-1022.

[5] Titov I, McDonald R, "Modeling online reviews with multi-grain topic models," *Proceedings of the 17th international conference on World Wide Web*. ACM, 2011,pp.111-120.

[6] Konishi T, Tezuka T, Kimura F, *et al*. "Estimating Aspects in Online Reviews Using Topic Model with 2-Level Learning," *Proceedings of the International MultiConference of Engineers and Computer Scientists*. 2012, 1.pp. 120-126

[7] Brody S, Elhadad N, "An unsupervised aspect-sentiment model for online reviews," Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012,pp. 804-812.

[8] Wang H, Lu Y, and Zhai C, "Latent aspect rating analysis on review text data: a rating regression approach," *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010,pp. 783-792.

[9] Wallach H M, "Topic modeling: beyond bag-of-words," *Proceedings of the 23rd international conference on Machine learning. ACM*, 2006,pp. 977-984.

[10] Griffiths T L, Steyvers M, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, 2004, 101(Suppl 1),pp. 5228-5235.

[11] Heinrich G. "*Parameter estimation for text analysis,"* Technical report, 2005.

[12] Blair-Goldensohn S, Hannan K, and McDonald R, "Building a sentiment summarizer for local service reviews," *WWW Workshop on NLP in the Information Explosion Era*. 2008: 14.

[13] Ferreira L, Jakob N, and Gurevych I. "A comparative study of feature extraction algorithms in customer reviews," *Semantic Computing, 2008 IEEE International Conference on. IEEE*, 2008,pp. 144-151.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014