Aspect Extraction from Product Reviews Using Category Hierarchy Information

Yinfei Yang
Redfin Inc.
Seattle, WA 98101 USA
yangyin7@gmail.com

Cen Chen
Singapore Management University
Singapore, 188065
cenchen.2012@phdis.smu.edu.sg

Minghui Qiu
Alibaba Group
Hangzhou, China 311121
minghuiqiu@gmail.com

Forrest Sheng Bao
University of Akron
Akron, OH 44325 USA
forrest.bao@gmail.com

Abstract

Aspect extraction is a task to abstract the common properties of objects from corpora discussing them, such as reviews of products. Recent work on aspect extraction is leveraging the hierarchical relationship between products and their categories. However, such effort focuses on the aspects of child categories but ignores those from parent categories. Hence, we propose an LDA-based generative topic model inducing the two-layer categorical information (CAT-LDA), to balance the aspects of both a parent category and its child categories. Our hypothesis is that child categories inherit aspects from parent categories, controlled by the hierarchy between them. Experimental results on 5 categories of Amazon.com products show that both common aspects of parent category and the individual aspects of subcategories can be extracted to align well with the common sense. We further evaluate the manually extracted aspects of 16 products, resulting in an average hit rate of 79.10%.

1 Introduction

E-commerce provides a whole new way for shopping that product reviews posted by some consumers can help others make their purchase decisions. One important task about online product review is to extract the properties of products, known as *aspects*. Aspect extraction has many applications, such as opinion mining (Liu, 2012; Liu et al., 2015), summerization (Bagheri et al., 2013; Hu and Liu, 2004), helpfulness prediction (Yang et al., 2016; Yang et al., 2015) and recommendation (Reschke et al., 2013; Jakob, 2011).

Statistical topic modeling, such as LDA (Blei et al., 2003) and its variants, has been shown to be successful for aspect extraction (Titov and McDonald, 2008; Zhao et al., 2010; Jo and Oh, 2011; Mukherjee and Liu, 2012; Moghaddam and Ester, 2013). Topic modeling clusters words based on their co-occurrences in sentences and documents to generate topics, each of which is a probabilistic distribution over words. Because words that co-occur are often about the same topic, which could talk about one aspect of a product, one or more aspects can be then associated with one or more topics. Earlier work of topic modeling is fully unsupervised while recently knowledge bases (KB) begin to be incorporated into semi-supervised schemes (Wang et al., 2014; Zhai et al., 2010; Chen et al., 2014).

However, existing approaches have limitations. First, the aspects usually become terms strongly associated with specific group of products (e.g., "multitouch" of touchscreen laptops), instead of the true ratable features of products (e.g., "battery life" for all laptops and even all portable electronic devices) (Titov and McDonald, 2008). Second, existing approaches require sufficient amount of corpora while many products do not have enough reviews, known as the *cold-start problem* (Moghaddam and Ester, 2013). For example, around 1/3 of the product categories used in our experiment from Amazon.com Review Dataset (McAuley and Leskovec, 2013) have less than 100 reviews. Third, current approaches do not provide a good balance between child category aspects and parent category aspects.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 675–680,
Valencia, Spain, April 3-7, 2017.

Therefore, we develop a new aspect extraction approach, called categorical LDA (**CAT-LDA**), by leveraging the hierarchy relationship between products. We hypothesize that reviews of each subcategory (e.g., gaming laptops) all contribute to the topics of its corresponding general category (e.g., laptops), but with different weights. As a result, aspects of a specific sub-category of products will be the combination of its unique aspects and the aspects from its parent (and thus shared with its siblings). This modeling also provides an approach to cold-starting problem by allowing aspects to be inherited from the parent category or transferred from sibling (sub-)categories.

Unlike most of the existing work modeling at the product item level, our model is based on the product category level. It can be easily extended to product item level by creating one node for each item and attaching them to the leaf nodes on the category hierarchy. Factorized LDA (FLDA) (Moghaddam and Ester, 2013) is based on the category level, but it only considers specific categories where all items in one category share a set of aspects. Our approach extends by modeling aspects in both the general and specific categories. Our model also relaxes the assumption in multi-grain LDA (MG-LDA) (Titov and McDonald, 2008) that only local topics contribute to product aspects, aligning better with common sense. Aspects at different layers are all related with each other through the product tree. For example, all portable electronic devices have a common aspect: battery life.

Empirical study is based on reviews from 5 general categories of Amazon.com Review Dataset (McAuley and Leskovec, 2013). The model we propose can generate human ratable product aspects from both general categories and sub-categories. We evaluate the extracted aspects for 16 product items of 9 categories against the annotations from (Hu and Liu, 2004; Ding et al., 2008; Liu et al., 2015). Promising experimental result shows 79% hit rate on manually annotated aspects.

2 Problem Formulation

In the context of the product aspect extraction, an *aspect* is an attribute or feature of a product item mentioned in reviews. Previous work of aspect extraction focuses on either an *aspect term* mentioned in review text or an *aspect category* which groups many aspect terms together (Zhai et al., 2010). Here we focus on the latter. However, we will show that our model is also able to detect aspect terms from an unseen text in Section 4.

In this paper, we propose a generative topic model with two layers of hierarchy: the *general categories* and the *sub-categories*. For example, "pocket watches" is a subcategory under the general category "watches". Product hierarchy information (also called *product tree*, Figure 2 as an example for "watches") can be extracted from online shopping websites, e.g., Amazon.com. For the sake of simplicity, we flatten the product tree into the two layers. General categories are at the top of product hierarchy and any category under it in the product hierarchy is its sub-category. It is still an open question to design a unified model to extract aspects by considering all the hierarchical layers.

Our goal is to identify the aspects of both general categories and sub-categories. We hypothesize that reviews under the same general category share some common aspects because of the similarity among them. But because of the difference among them, each subcategory has its unique aspects.

3 Methodology

According to our hypothesis, when composing a review, a consumer considers aspects of both the general category and the subcategory that the product belongs to. Such generative process can be represented in the graphical model as in Figure 1. We refer to "aspect" as "topic" in the context of topic modeling.

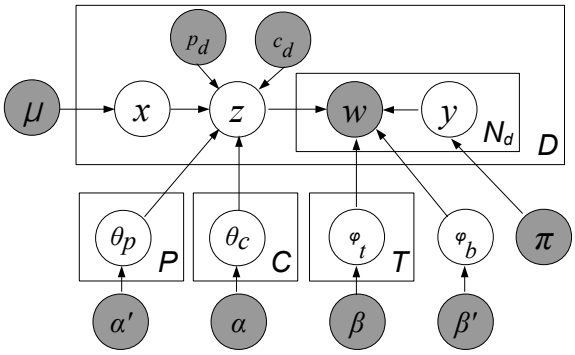

Figure 1: A graphical model representation of review generation.

Denote P as the set of general categories and C the set of sub-categories. Each general cate-

gory $p \in P$ has a topic distribution θ_p while each sub-category $c \in C$ has a topic distribution θ_c. When generating a sentence, a topic distribution is picked first using a switch x following Bernoulli distribution μ. Like in standard topic modeling, each topic t is a distribution over words, denoted as φ_t. Further, there is a set of background words whose distribution is denoted as φ_b. To choose between background words and topic words, we assume another switch y following Bernoulli distribution π.

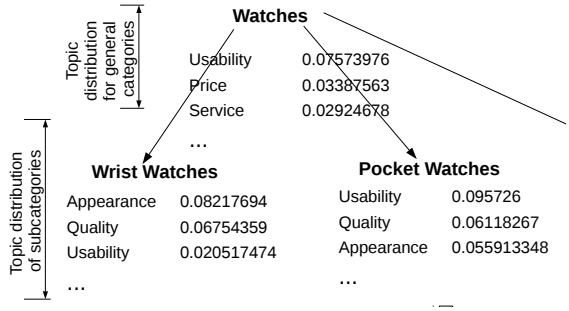

Figure 2: An example illustrating how reviews are generated.

When a sentence is generated, given its subcategory c and its general category p, we first sample a value for switching x based on μ. Let $\theta = \theta_c$ (e.g., "wrist watches" or "pocket watches" in Figure 2) if $x = 0$ (i.e., picking the topic of a sub-category), otherwise $\theta = \theta_p$ (e.g., "watches" in Figure 2) (i.e., picking the topic of a general category). A topic z is chosen based on the topic distribution θ. For each word position in the sentence, first sample a value for switching y based on π and then pick the word based on the word distribution φ_t of the topic z if $y = 0$, or from background word distribution φ_b otherwise. Figure 2 illustrates the generative process using watches as an example, showing top 3 aspects and their probabilities.

All distributions θ_c, θ_p, φ_t, φ_b are generated from Dirichlet priors with hyperparameters α, α', β, and β', respectively. The generation process is:

1. For each general category $p \in P$, choose $\theta_p \sim \text{Dir}(\alpha')$
2. For each subcategory $c \in C$, choose $\theta_c \sim \text{Dir}(\alpha)$
3. For each aspect $t \in T$, choose $\varphi_t \sim \text{Dir}(\beta)$
4. For background words, choose $\varphi_b \sim \text{Dir}(\beta')$
5. For each sentence (a document) $d \in D$,
   (a) Get its specific sub-category c and general category p from meta data
   (b) Choose a switch $x_d \sim \text{Bernoulli}(\mu)$
   (c) Choose an aspect $z_d \sim \text{Multi}(\theta_c)$ if $x_d = 0$, otherwise $z_d \sim \text{Multi}(\theta_p)$
   (d) For each word $n \in \{1, 2, \ldots, N_d\}$,
      i. Choose a balance $y_{d,n} \sim \text{Bernoulli}(\pi)$
      ii. If $y_{d,n} = 1$, choose a topic word $w_{d,n} \sim \text{Multi}(\varphi_{z_d})$; else choose a background word $w_{d,n} \sim \text{Multi}(\varphi_b)$.

where N_d means the number of words in document d, "Dir" refers to "Dirichlet", and "Multi" refers to "Multinomial". Each multinomial distribution is governed by some symmetric Dirichlet distribution. We use Gibbs sampling to perform model inference and present the sampling formulas as follows.

Let τ be the set of hyperparameters $\{\alpha, \alpha', \beta, \beta', \mu, \pi\}$, c, p be the sub-category and general category of document d's n-th aspect. We collapse out all the θ_c, θ_p, φ_t, and φ_b, and jointly sample switch x_d and aspect label z_d as follows:

$$p(z_d = t, x_d = 0 \mid Z_{\neg d}, Y, W, \tau)$$
$$\propto \frac{n_{x=0} + \mu - 1}{n_{\cdot} + 2\mu - 1} \cdot \frac{n^t_{x=0,c} + \alpha - 1}{n^{\cdot}_{x=0,c} + T\alpha - 1} \cdot \frac{\prod_{w=1}^{V} \prod_{p=1}^{n^w_d} (n^{t,y=1}_w + \beta - p)}{\prod_{q=1}^{n^{\cdot}_d} (n^{t,y=1}_{\cdot} + V\beta - q)},$$
$$p(z_d = t, x_d = 1 \mid Z_{\neg d}, Y, W, \tau)$$
$$\propto \frac{n_{x=1} + \mu - 1}{n_{\cdot} + 2\mu - 1} \cdot \frac{n^t_{x=1,p} + \alpha' - 1}{n^{\cdot}_{x=1,p} + T\alpha' - 1} \cdot \frac{\prod_{w=1}^{V} \prod_{p=1}^{n^w_d} (n^{t,y=1}_w + \beta - p)}{\prod_{q=1}^{n^{\cdot}_d} (n^{t,y=1}_{\cdot} + V\beta - q)}.$$

where $n^t_{x=0,c}$ is the number of times topic t and sub-category c co-occur, and $n^t_{x=1,p}$ is the number of times topic t and general category p co-occur.

Similarly, we sample $y_{d,n}$ as follows:

$$p(y_{d,n} = y \mid Y_{d,\neg n}, Z, W, \tau)$$
$$\propto \frac{n_y + \pi - 1}{n_{\cdot} + 2\pi - 1} \cdot \Big[\frac{n^{t,y=1}_w + \beta - 1}{n^{t,y=1}_{\cdot} + V\beta - 1}\Big]^{y=1} \cdot \Big[\frac{n^{y=0}_w + \beta' - 1}{n^{y=0}_{\cdot} + V\beta' - 1}\Big]^{y=0}.$$

4 Experiment

Reviews from 5 categories (details in Table 1) of Amazon.com Review Dataset (McAuley and Leskovec, 2013) are used as the corpora. A total of 200 topics are built.

Table 1: The 5 categories used to model topics

General category	# of sub-categories	# of reviews
baby products	226	184,887
watches	10	68,356
software	171	95,084
cellphones	33	78,930
electronics	674	1,241,778

4.1 Qualitative Results

We select top topics at different levels and manually examine if they can be aligned with some

Table 2: Top topics and topic words for each General Category. Labels are manually assigned.

Category	Label	Top Words
baby	Value	money, worth, waste, time, buy, product, price, save, spend, good...
	Shipping	great, product, arrived, fast, quality, shipping, easy, advertised, received, delivery...
	Return	amazon, return, shipping, back, days, received, item, order, ordered, refund...
watches	Wrist	watch, band, wrist, face, watches, easy, strap, size, read, wear ...
	Return	amazon, return, shipping, back, days, received, item, order, ordered, refund...
	Quality	quality, made, good, plastic, cheap, solid, sturdy, feels, product, construction...
software	Product	software, version, program, product, cd, computer, buy, easy, upgrade, install...
	Support	support, tech, customer, call, phone, service, called, problem, hours, email...
	Install	easy, manual, instructions, set, user, simple, install, setup, read, software ...
cellphones	Headset	headset, headsets, bluetooth, hear, sound, quality, volume, ear, noise, phone...
	Review	reviews, review, product, read, bad, problems, good, problem, write, negative...
	Case	case, phone, clip, screen, belt, cover, fit, fits, plastic, leather...
electronics	Value	money, worth, waste, time, buy, product, price, save, spend, good...
	Return	amazon, return, shipping, back, days, received, item, order, ordered, refund...
	Shipping	great, product, arrived, fast, quality, shipping, easy, advertised, received, delivery...

certain aspects. Because the top ranked topics are equivalent to the topics mentioned the most in reviews, we can treat these topics as the most important aspects. For better representation, we also manually assign an "aspect" label to each topic.

Top words for the top topics discovered in each general category are presented in Table 2 in the form of one topic per line, along with the top ranked words in this topic. For space sake, only three topics are presented. They align well with the product aspects in our common sense.

For example, Value is the most cared aspect of baby product buyers, followed by Service and Return. The electronics products have the same highest ranked aspects, but in a different order. Unlike other categories, the top aspects for Software are Product, Support and Install, which are unique aspects of software in our common sense.

Table 3 shows the top five topics and top words among all categories. Not surprisingly, Value, Return and Shipping are still the most important aspects for customers who shop online. Review, basically "the reviews from other customers", is also mentioned frequently, indicating that customers are indeed influenced by the reviews of others. In the end, people like to talk about their Experience and compare to that with other retailers, local or online.

Table 3: Top topics and topic words across all categories. Labels are manually assigned.

Label	Top Words
Value	money, worth, waste, time, buy...
Return	amazon, return, shipping, back, days...
Shipping	great, product, arrived, quality, fast...
Review	reviews, review, product, read, bad...
Experience	price, amazon, shipping, store, deal...

Table 4: Top topics and topic words for Laptops. Labels are manually assigned.

Label	Top Words
Spec	ram, memory, computer, card, video...
Design	mouse, keyboard, keys, buttons, wireless...
System	version, windows, mac, xp, os...
Warranty	warranty, back, service, unit, repair...
Screen	screen, picture, monitor, color, bright...

Lastly, we are interested in top topics for specific categories. Due to space limit, we pick Laptop Computers to study (Table 4). Quite unlike topics for general category, the top topics for Laptops are very product related: Spec, Design, System, Warranty and Screen.

4.2 Quantitative Results

We then quantitatively study whether our model can really extract aspects. The ground truth is the sentence-level manual aspect annotations in a combined dataset from (Hu and Liu, 2004; Ding et al., 2008; Liu et al., 2015), which contains 10,993 reviews of 17 products in total. The aspects are annotated at sentence level. Among them, we select 16 products that can be linked to the 5 general categories used to train our model above. The 16 products belong to 9 categories (Table 5). Note that not all sentences are annotated, we only predict the sentences with human annotations. For comparison, MG-LDA (Titov and McDonald, 2008) is used as the baseline.

We first attach each product to its closest category in the category hierarchy. For each sentence with manual aspect annotations, the model described above is used to find its most like topic. Then we select 3 words from the sentence with the highest probability under the detected topic as

highlighted words, hoping that highlighted words can cover the aspect terms annotated manually. However, the manual annotations can also involve words not in the sentence. So we also include the top 3 topic words of the detected topic because they are the best words to describe the topic.

Table 5: Hit rates of aspect by topic work

Category	# of products	# of sentences	CAT-LDA	MG-LDA
Digital camera	4	697	85.7%	65.7%
DVD player	1	344	79.1%	72.1%
MP3 player	3	1,356	74.7%	60.3%
Audio speaker	1	301	89.7%	70.9%
PC monitor	1	239	91.2%	72.3%
Network router	2	437	79.0%	69.1%
Cell phone	2	629	80.8%	72.8%
Diaper champ	1	212	66.0%	60.8%
Anti-virus software	1	210	68.6%	60.0%
Average	–	–	79.1%	67.1%

We say a "hit" if the highlighted words and top 3 topic words of a sentence cover all manually annotated aspect words, and a "miss" otherwise. For example, given a camera review *Also as someone who at least knows a little bit about the technical work of taking a photo i really miss having manual controls.* Words *manual controls* are annotated as aspect terms. The highlighted words extracted by CAT-LDA are *photo*, *manual* and *controls*, and the topic words are *control*, *controls*, *remote*. It is a "hit" because the aspect terms are covered by highlighted words and topic words. The hit rates of different products are given in Table 5. To be fair, sentences used for the quantitative test are not used to train the topic models.

Because MG-LDA is not originally designed for extracting aspects for general categories, we train one MG-LDA model for each category in Table 5 to avoid introducing a disadvantage for MG-LDA [1]. Similar to above, we first find the closest category for each product in the category hierarchy and then train a model on all reviews of this category.

The result of CAT-LDA is very promising, with an average hit rate of 79.10% among all 9 categories of products. Physical products of computer or electronics type have very high hit rates, with the highest 91.21% for PC monitors. The low hit rates of diaper champ and software are due to the lack of components, especially descriptive ones, and their limited functionality. CAT-LDA leads MG-LDA in all of 9 categories of products with an average hit rate improvement of 12%.

The results can be further improved if we consider synonyms words of aspect terms or adding more features like Part-of-Speech tags and dependence rules (Hu and Liu, 2004; Yu et al., 2011). Because it is not the main focus of this paper, we leave it as future work.

[1] We have tried training one MG-LDA model for each of the 5 general categories but the results for MG-LDA are not as good.

5 Conclusion

In this paper we propose a generative model for aspect extraction leveraging product category hierarchy. Our hypothesis is that any product's aspects are a mixture of aspects from its parent category and aspects unique to itself. Topic models built in this way can successfully balances the aspects of a product itself and its parent category. Experimental results show 79% hit rate on manually annotated aspect terms of 16 products covering 9 categories.

References

Ayoub Bagheri, Mohamad Saraee, and Franciska de Jong. 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, 52:201–213, November.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 347–358, Baltimore, Maryland, June. Association for Computational Linguistics.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. *Proceedings of the international conference on Web search and web data mining - WSDM '08*, page 231.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, 04:168.

Niklas Jakob. 2011. *Extracting Opinion Targets from User-Generated Discourse with an Application to Recommendation Systems*. Ph.D. thesis, Technische Universität.

Yohan Jo and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA. ACM.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1291–1297. AAAI Press.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

J. McAuley and J. Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, pages 165–172.

Samaneh Moghaddam and Martin Ester. 2013. The flda model for aspect-based opinion mining: Addressing the cold start problem. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 909–918, New York, NY, USA. ACM.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 339–348, Jeju Island, Korea, July. Association for Computational Linguistics.

Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–504, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 111–120, New York, NY, USA. ACM.

Tao Wang, Yi Cai, Ho-fung Leung, Raymond Y.K. Lau, Qing Li, and Huaqing Min. 2014. Product aspect extraction supervised with online domain knowledge. *Knowledge-Based Systems*, 71:86–100, November.

Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 38–44, Beijing, China, July. Association for Computational Linguistics.

Yinfei Yang, Cen Chen, and Forrest Sheng Bao. 2016. Aspect-based helpfulness prediction for online product reviews. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 836–843, Nov.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1496–1505, Portland, Oregon, USA, June. Association for Computational Linguistics.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2010. Grouping product features using semi-supervised learning with soft-constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1272–1280, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65, Stroudsburg, PA, USA. Association for Computational Linguistics.