Summer 6-30-2018

Fine-grained Aspect Extraction for Online Reviews of E-commerce Products Based on Semi-supervised Learning

Huosong Xia
School of Management, Wuhan Textile University, Wuhan, 430073, China

Yitai Yang
School of Management, Wuhan Textile University, Wuhan, 430073, China

Follow this and additional works at: http://aisel.aisnet.org/whiceb2018

Fine-grained Aspect Extraction for Online Reviews of E-commerce Products Based on Semi-supervised Learning

Huosong Xia[*]*, Yitai Yang*

School of Management, Wuhan Textile University, Wuhan, 430073, China

Abstract: The accuracy of online review mining for e-commerce products is of great value to customer and product matching portrait. Mining the fine-grained aspect in reviews is a key indicator. It can better analyze the emotion tendency of online reviews and understand the advantages and disadvantages of evaluation objects. In this paper, we propose a semi-supervised learning method to extract product aspects and description of aspects. Specifically, we firstly construct word vector space model of large scale reviews with deep learning, then get the list of similar words based on the model. Finally, the fine-grained aspect sets are obtained by classification algorithm. The results of the study show that the efficiency of fine-grained extraction is improved by using semi-supervised method.

Keywords: Fine-grained aspect, car reviews, aspect extraction, deep learning

1. INTRODUCTION

Online reviews are important information sources for business data analysis in e-commerce matching portrait between customer and product. Mining sentiment information in online reviews is of great value for e-commerce. Aspect extraction is a key part of online review mining. Extracting the aspects of users' concern in online reviews can be more effective in mining useful information. As the amount of online review data is growing, it is possible to understand the global view of the evaluation object by analyzing the more finer attribute sentiment. In order to solve this problem, a PageRank algorithm to extract product features has been proposed in Yan et al.[1], a method with the syntactic parsing and K-means to extract fine-grained aspect has been proposed in Qiu et al.[2], a LDA method has been proposed to get fine-grained aspect in Jia et al.[3]. But these studies can only extract the main aspects of evaluating objects for the limitations of adopting methods and the stage of technological development.

Owing to the development and application effect of deep learning which has been generally consented by both academia and industry, this paper proposes a semi-supervised method to get fine-grained aspects in online reviews with deep learning, which aiming at solving the global view and efficiency of this problem. Our method is experimented on car reviews, and compared with the new method in Zhou and Zhang[4]. The results show that this method can obtain more fine-grained attributes while ensuring higher efficiency. The remaining main structures of the paper are as follows: the second part is related research and theoretical methods, the third part is the fine-grained aspect extraction method of Chinese online review, the fourth part is the experimental results and discussion, the last part is the discussion and conclusion.

2. RELATED WORK

Fine-grained aspect extraction is an important part of online review mining. The methods of extracting fine-grained aspect are divided into two categories: fine-grained aspect extraction based on ontology and fine-grained aspect extraction based on machine learning[5]. The fine-grained extraction method based on ontology refers to extracting fine-grained based on setting up the concept of evaluation object aspect and the

relationship between object aspect and itself[6]. Lau et al. construct a fuzzy domain ontology tree with emotional dictionary and the evaluation object[7]. Lau et al. propose a method using LDA and Gibbs to analysis fine-grained aspect emotion[8]. Tang et al propose a method to get fine-grained aspects of product with frequent item algorithm, establish product feature ontology to get feature emotion word pairs, and then obtain implicit product features based on feature emotion word pair[9].

The fine-grained extraction method based on ontology can only get the main evaluation object attributes, and it is difficult to excavate more fine-grained attributes. Therefore, more researches extract fine-grained attributes using machine learning. Fine-grained attribute extraction methods based on machine learning can be divided into two categories, one is traditional machine learning method, the other is deep learning. In the traditional method of machine learning, extracting fine-grained aspects are mainly based on the rule based learning method. Some methods use building dictionary to get fine-grained aspects, some methods use LDA, CRF and other models to get fine-grained aspects according to contextual information, and some of them use clustering to get fine-grained attribute [10, 11]. Amplayo et al. propose a model named EBTM to extract fine-grained aspects in short text[12]. Poria constructs aspect extraction rules with domain knowledge and syntactic dependency to obtain fine-grained aspects[13]. In contrast with supervised methods, unsupervised methods mainly use clustering methods to extract product attributes.

As the number of online reviews is increasing, the available online reviews data is getting bigger and bigger. The results of extracting fine-grained aspect with deep learning in the large data set are often better than the traditional machine learning methods. Zhou et al. obtain word vectors with deep learning, and then obtain fine-grained aspect based on the AP clustering algorithm and domain knowledge. The method proposed by Zhou et al. has more fine-grained aspects than traditional machine learning methods. But this method still needs after clustering while manually getting obtain fine-grained aspect. Therefore, this paper proposes a semi-supervised method to extract fine-grained aspect. First, we build the word vector model of online reviews with deep learning, and then get the list of similar words based on the word vector model. Finally, we use the classification algorithm to get the fine-grained aspect set automatically.

3. FINE-GRAINED ASPECT EXTRACTION IN CHINESE ONLINE REVIEWS

In this paper, a semi-supervised method is proposed to extract the aspect of the user's attention in the online reviews. Firstly, we construct a word vector model of online reviews with neural network, and then get the list of similar words of the candidate aspect words based on the model. Finally, the words in the list are divided into three categories with a classification algorithm. The three categories are the same aspect words, the fine-grained words and the emotion words. The same words and the fine-grained words are the final fine-grained aspect set. As shown in Fig. 1, the first step is to preprocess the online reviews, it is followed by a neural network model, which is a training word vector model with skip-gram. The second step takes the key aspect words as the seed words, and obtains the similar words list of the key aspect words according to the word vector model. Finally, the words in the list are divided into three categories with a classification algorithm. The three categories are the same aspect words, the fine-grained words and the emotion words. The meaning of words in the same aspect words is the same as key aspect word. The fine-grained words contain the fine-grained aspect words of the key aspect word in the user reviews. The emotion words include the emotion words on the key aspect of the evaluation objects and their fine-grained aspect in reviews.
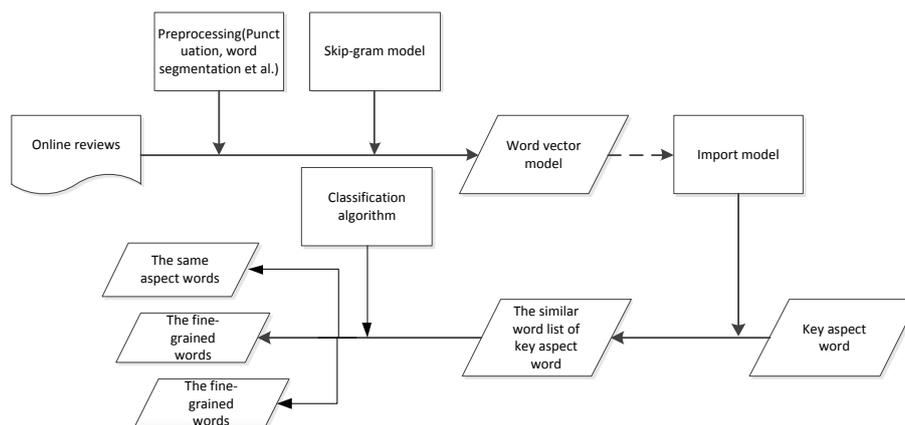
Figure 1. Fine-grained attribute extraction method

3.1 Word vector model

There are many computing methods for text semantic similarity. The way of word vector is to map words as features into word vectors, and then calculate the similarity between words and words by calculating the distance between vectors and vectors. This paper trains word vector model of online reviews by skip-gram neural network. The skip-gram model is a three layer neural networks, it can predict the probability of its adjacent words by each word itself, as shown in Fig. 2.
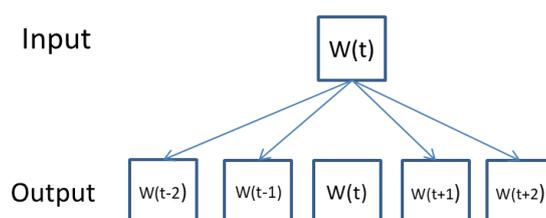

Figure 2. Skip-gram model

Skip-gram model takes the word as features, mines deeper features to get word vector of words by mapping features to K dimensional vector space, and uses cosine similarity to calculate the similarity between the words, the approach can be expressed via Eq.(1). The greater result of $\cos(W_i, W_j)$, the higher the similarity between words. In this paper, the skip-gram model in Word2Vec is used to train online reviews to obtain the word vector model, and each candidate aspect word can get the list of similar words based on the word vector model.

$$\cos(W_i, W_j) = \frac{\sum_{k=1}^{N}(W_i^k \times W_j^k)}{\sqrt{\sum_{k=1}^{N}(W_i^k)^2} \times \sqrt{\sum_{k=1}^{N}(W_j^k)^2}} \tag{1}$$

Here, W_i^k and W_j^k denote the weight of the i-th and j-th word vector in k dimensional respectively. In addition, N denotes the number of word vector dimensions.

3.2 The fine-grained aspect classification algorithm

There are three categories words in the list of similar words. The three categories are the same aspect words, the fine-grained words and the emotion words. The meaning of words in the same aspect words are the same as key aspect words. The fine-grained words contain the fine-grained aspect words of the key aspect word in the user reviews. The emotion words include the emotion words on the key aspect of the evaluation objects and their fine-grained aspect in reviews.

Many online reviews have different appellations for the same aspect because of the high degree of freedom

and randomness. Such as word "内饰" has written in "内室", this phenomenon of error writing is very common in online reviews. This paper proposes a way to mining the words with same meaning to key aspect word in review. First, words in the list of similar words are transformed into the phonetic alphabet, this is followed by calculating weight of each word based on the Levenshtein distance. Then the words is classified as the same aspect word if the weight of word is less than the threshold. The Levashtein distance between two words is the minimum number of single-character edits required to change one word into the other one, the edit operations include insertions, deletions and substitutions. The same aspect words algorithm as shown in Table 1.

Table 1. Classification algorithm for the same aspect words

Name: getCateA()

Input: simList, keyword, L   //*simList is the similar word list of keyword*, *L is the threshold of distance*

Output: cateA   //*cateA includes words with the same mean of keyword*

Init: cateA=∅

Begin:

    pyKeyWord = keyword.pinyin()   //*get pinyin of keyword*

    for word in simList:

        pyWord = word.getpinyin()

        simDis = Levenshtein.distance(pyword,pyKeyWord)

        if simDis < L:

            cateA.add(word)

    return cateA

End

Another part of the fine-grained aspect classification algorithm is to obtain the fine-grained words and the emotion words. The idea of the algorithm in this paper is as follows: The first step is to construct an emotional word dictionary based on the emotional words of HowNet and the emotional vocabulary ontology library of Dalian University of Technology[16]. The second step is to classify words into the emotion words if the words appeared in the emotional word dictionary. The third step, if the word do not appear in the emotional word dictionary, the word is used as the key word to obtain its list of similar words. The score of each word in the list is calculated, the score of word is the similarity weight between the word and it's key aspect word if word is not an emotional word, otherwise the score will be opposite, then score of the word is calculated by sum weight of its list of similar words. Word is classified in the fine-grained words if score is negative, otherwise classified in the emotional words if score is positive. The algorithm to obtain the fine-grained words and the emotional words as shown in Table 2 and Table 3.

Table 2. Recognition algorithm for emotional words

Name:get_emotion()

Input model, keyword, score, deepFlag, emotionDict. //*deepFlag is the depth hreshold of recursion, emotionDict is a directory of*
    //*emotional words. model is the word vector model. score is the similar score of keyword*

Output score //*return 1 if keyword is an emotional word otherwise -1 if not*

Begin

    deepFlag = deepFlag-1

    if deepFlag < 0:

```
                return -1*score
        if keyword in emotionDict:
                return 1*score
        else:
                simList = model.most_similar(keyword,num=10)//get 10 words most similar to keyword and their scores
                for word in simList:
                        score = score+get_metion(word,model,emotionDict, score, deepFlag)
                if score <= 0:
                        return -1*score
                else:
                        emotionDict.add(word)
                        return 1*score
```

Table 3. Classification algorithm for the fine-grained words and the emotional words

Name: get_cateBC()

Input simList, model, keyword, emotionDict, deepFlag

Output cateB, cateC   *//cateB includes fine-grained aspect words of keyword, cateC includes emotional words*

Init: cateB=∅, cateC=∅

Begin

```
        For word in simList:
                If get_motion(word,model, emotionDict, word.score, deepFlag)
                        cateB.add(word)
                else
                        cateC.add(word)
        return cateB, cateC
```

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Experimental data

The datasets used in this study were obtained from www.autohome.com.cn. We extracted 815557 Chinese online reviews from the website, and divided the reviews into sentences by punctuation because the reviews are too long. We get 50505086 sentences after preprocessing. Part of data as shown in Table 4

Table 4. Sample of car reviews

Number	Review
1	四门具备一定实用性
2	小众车满足装 X 欲望
3	油耗说实在的都买这排量的车了
4	座椅的包裹性强
5	这种车没有性价比一说

4.2 Experimental procedure and results

Reviews are divided into sentences by punctuation in our method. It is followed by segmentation with Jieba, a tool to separate a Chinese sentence into words. And then the skip-gram model in Word2Vec is used to train online reviews to obtain the word vector model. In this paper, the results of fine-grained aspect words are extracted from two key aspect words, "内饰"(Interior) and "外观"(Appearance), the fine-grained attribute extraction method proposed by Zhou Qingqing et al. is used as a contrast[4]. Part results of fine-grained aspects are shown in Table 5.

Table 5. Part results of fine-grained aspects

Aspect	Method	Fine-grained aspect	Emotion word
Interior	Our method	'内设', '内适', '黑内', '总体设计', '用料', '棕黑', '风格', '做工', '配色', ', '料子', '前中控'…	'大气', '简简单单', '简洁', '时尚', '素雅', '赖看', '新潮', '不拉风', '前卫', '大大方方', '上档次'
	Compared method	'内设', '内适', '木纹', '黑内', '总体设计', '用料','做工', '配色', ', '料子', '材料', '前中控'	∅
Appearance	Our method	'外怪', '外感', '外款', '屁股', '很靓', '前脸', '红颜色', '车脸', '咖版', '外表', '大胡子', '侧身' …	'犀利', '大气', '帅气', '可爱', '不娘', '时尚', '赖看', '霸气', '飘亮', '大气磅礴', '高大威武'
	Compared method	'外怪', '外感', '外款', '屁股', '很靓', '前脸', '红颜色', '车脸', '咖版', '外表', '大胡子', '侧身'…	∅

4.3 Experimental analysis

As the data set used in this article is large, it is costly to annotate for all online reviews, so this article builds the aspect set as follows: We search for twenty videos which introduce or recommend cars, and treat the description words of a car in the video as a fine-grained aspect when it is used to introduce the interior and appearance details. In the final result, the fine-grained aspect set of the interior contains 47 attributes, with 39 attributes in the appearance, and part result of aspects as shown in Table 6.

Table 6. Part result of aspects of interior and appearance

Aspect	Fine-grained aspect
Interior	'前中控','材料', '造型', '质地', '工作台', '中控台', '布局'
Appearance	'屁股', '前脸', '侧身', '侧脸', '外形', '流线形','线条'

This paper uses recall score, accuracy score and F1 value to evaluate the results of fine-grained properties, the approach can be expressed via Eqs. (2)–(4):

$$P = \frac{TP}{TP+FP} \tag{2}$$

$$R = \frac{TP}{TP+FN} \tag{3}$$

$$F_1 = \frac{2P*R}{P+R} \tag{4}$$

Here, *TP* denotes the number of extracted aspects correctly, *FP* denotes the number of extracted aspects inaccurately, *FN* denotes the number of non-extracted aspects, P denotes accuracy score, R denotes recall scores. The evaluation results are as shown in Table 7.

Table 7. The evaluation results on car reviews

	Accuracy		Recall		F_1	
	Our method	Compared method	Our method	Compared method	Our method	Compared method
Interior	78.6%	81.1%	93.6%	91.4%	85.4%	85.9%
Appearance	77.8%	83.3%	89.7%	89.7%	83.3%	86.3%

In Table 7, the accuracy and F_1 of our method is lower than the compared method. The reason is that our method obtains fine-grained aspect automatically by a classification algorithm, when the fine-grained aspect sets are obtained, but the compared method uses the artificially selected method to filter the noise vocabulary and reduces the number of the aspect of the error extraction. The recall scores in Table 7 show that our method is better than the compared method, the result demonstrates that our method can be more comprehensive to extract the fine-grained aspect from online reviews. In contrast with compared method, the results of two methods are in the same level, but our method extracts fine-grained aspect with more efficiency by a semi-supervised method.

5. CONCLUSIONS

With a heavier demand on global view of e-commerce matching portrait between customer and product, this paper proposes a method to extract fine-grained with technical background of the development of deep learning. Our method improved the problem of efficiency and granulation in the existing research. Firstly we construct a word vector model of online reviews with neural network, and then get the list of similar words of the candidate aspect words based on the model. Finally, the words in the list are divided into three categories with a classification algorithm. The three categories are the same aspect words, the fine-grained words and the emotion words. The same words and the fine-grained words are the final fine-grained aspect set. In contrast with the traditional method based on seed word or LDA, this method can obtain more meaningful fine-grained attributes. Our method improves the efficiency of fine-grained extraction with a semi-supervised approach, In contrast with a similar new methods in literature[4].

However, this method also has some shortcomings. Although we can get more fine-grained aspects, some of the emotional words are divided into the final fine-grained aspect set because of noise. The future work will reduce noise in fine-grained aspect extraction with unsupervised or semi-supervised algorithms. At the same time, it can expand to the more common e-commerce product reviews mining in the application area, promote the pixel of the global view of matching portrait between customer and product, and achieve the precise mining.

REFERENCES

[1]   Yan Z, Xing M, Zhang D, Ma B. (2015). Exprs: an extended pagerank method for product feature extraction from online consumer reviews. Information & Management, 52(7), 850-858.

[2]   Qiu Y, Ni X, Shao L. (2015). Research on extracting method of commodities implicit opinion targets. Computer Engineering and Applications, 51(19):114-118(in Chinese).

[3]   Jia W, Zhang H, Yang C, Zhao X, Li Bo. (2016). User sentiment model oriented to product attribute. Journal of Computer Applications, 36(1):175-180(in Chinese)

[4]   Zhou Q, Zhang C. (2017). Fine-grained aspect extraction from online customer reviews. Journal of the China Society for Scientific and Technical Information. 36(5):484-493(in Chinese)

[5]   Tang X, Liu G. (2017). Research review on fine-grained sentiment analysis. Library and Information Service, 61(5):132-140(in Chinese)

[6]   Silveira Chaves M, Laurel A., Sacramento N, Drebes Pedron C. (2014). Fine-grained analysis of aspects, sentiments and types of attitudes in restaurant reviews. Encontros Cient ficos - Tourism & Management Studies, 10(1): 66-72.

[7]   Liu L, Nie X, Wang H. (2012). Toward a fuzzy domain sentiment ontology tree for sentiment analysis. International Congress on Image and Signal Processing, IEEE, 1620-1624.

[8]   Lau R Y K, Li C, Liao S S Y. (2014). Social analytics: learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decision Support Systems, 65(5):80-94.

[9]   Tang X, Lan Y. (2016). Sentiment analysis of microblog product reviews based on feature ontology. Library and Information Service, 60(16):121-127,136(in Chinese)

[10]   S änger M, Leser U, Klinger R. (2017). Fine-Grained Opinion Mining from Mobile App Reviews with Word Embedding Features. Natural Language Processing and Information Systems, pp3-14.

[11]   Chen Z Y, Mukherjee A, Liu B. (2014). Aspect extraction with automated prior knowledge learning.   Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 347-358.

[12]   Amplayo R K, Song M. (2017). An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews. Data & Knowledge Engineering, 110(3):54-57.

[13]   Poria S, Cambria E, Ku L. (2014). A rule-based approach to aspect extraction from product reviews. Proceedings of the 2nd Workshop on Natural Language Processing for Social Media, 28-37

[14]   Suleman K, Vechtomova O. (2015). Discovering aspects of online consumer reviews. Journal of Information Science, 42(4):492-506.

[15]   Zhao L, Huang M, Sun J, Luo H, Yang X, Zhu X. (2015). Sentiment extraction by leveraging aspect-opinion association structure, Proceedings of the 24m ACM International Conference on Information and Knowledge Management. New York: ACM. 343-352

[16]   Xu L, Lin H, Pan Y, Ren H, Chen J. (2008). Constructing the Affective Lexicon Ontology, Journal of the China Society for Scientific and Technical Information. 27(2):180-185(in Chinesse)