Journal of Computing and Electronic Information Management

ISSN: 2413-1660

Research on implicit Feature Extraction Algorithm of online comments

Tinting Zhang

Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: As the Internet becomes more and more closely connected with our daily life, opinion mining for users' online comments in the field of e-commerce has become a hot research direction. At present, most of the online reviews of feature extraction for explicit characteristics, but the online comments this essay the problems tend to be colloquial, many sentences lack of a clear characteristics, to extract implicit characteristics can be used in a more comprehensive analysis users online reviews to emotional tendencies in order to provide users with the features of products purchase reference, to help businesses to improve product performance, can also be applied to features in the expansion of recommendation algorithm to recommend the result more accurate, etc. This paper presents a simple and effective method to select explicit sentences as training set and implicit sentences as test set. Antonyms are introduced to expand the dictionary of collocation rules, and the mapping relation of the dictionary of collocation rules is used to find implicit features. Compared with the traditional method, the recall rate and accuracy of this paper have been increased by 11.5% and 2.2% respectively. The experimental results are good. Moreover, this paper's method is simple and easy to implement with less manual intervention, which is of certain practical value.

Keywords: Opinion mining; Implicit features; Explicit features; Feature recognition; Online review.

1. Introduction

With the development of e-commerce, the user feedback information of online reviews quantity rapid growth, due to the large increase in e-commerce enterprises and users, potential users in decision-making when it is difficult to see the feedback information, producers need effective methods to monitor the feedback, the feedback reflects the consumer opinion of a product or service and attitude, the feedback has the very high use value.

One reason is that opinions and emotions expressed in comments can have an impact on other customers' buying intentions;

Another reason is that it is convenient for merchants to improve product quality or service, which can improve customer satisfaction, which is also a powerful basis for developing and selling new products.

Compared to the traditional market research, the use of acquisition of goods online comments after analysis can obtain more extensive user opinions, can also be convenient to track the evolution of opinion over time, can help customers recommend and allow them to target sales, thus, product features extracted from user online reviews is a child of the research field of opinion mining [1].Through a preliminary and rough analysis of the critical text, it is found that there are a large number of implicit feature sentences in the critical text, that is, there are no specific feature words in the review. The concept of implicit feature was first proposed by Su et al[2], Explicit features in display comments refer to words that appear explicitly in a sentence, such as " The price is not expensive ", where "Price" is an explicit feature; Implicit features refer to words that do not actually appear in the sentence but are implied by some opinion words, such as " The phone looks beautiful ", in which the feature word " beautiful " indicates the " appearance ", which does not actually appear in the sentence, but is implied by "beautiful". Most of the current studies only focus on explicit features [3-7]，If the implicit features mentioned above can be more accurately identified, it will be of great help to more comprehensively identify the emotion recognition in user comments and for the follow-up.

At present, most studies on implicit feature extraction are based on statistical or semantic algorithms to construct feature-opinion word collocation pairs, and then match feature words for opinion words in implicit comment sentences based on the collocation relation between feature words and opinion words [8].This method was first proposed by Liu et al [9]. In their proposed method, the mapping from opinion words to feature words is established by establishing association rules of feature words and opinion words, and then the implicit features are identified on the basis of this mapping relationship. Xu [10] An extended topic model is proposed to construct the displayed topic model and extract implicit features by combining prior knowledge. Zhang li [11] proposed a dictionary based on the dictionary and combined with the multi-strategy implicit feature extraction algorithm to extract implicit features. The algorithm is a dictionary constructed on the basis of multi-word words to reduce opinion word clusters and feature words of common words in the field.

The above scholars have made some achievements in the research of implicit feature recognition, but there are still some problems. Most of the current studies rely on a large number of manual annotations in the early stage. This method is time-consuming

and labouring, and there are still some areas that can be improved. In addition, how to build a more comprehensive and effective rule dictionary plays a very important role in improving recall rate. Based on these problems, this article embarks from the text preprocessing step, and considering the particularity of the implicit feature extraction, which need to be according to characteristics of the sentence as the training set, does not have characteristics as a test set, this paper puts forward a simple and effective characteristics of original data set is divided into contain explicit test set and test set contains no explicit characteristics. In the process of constructing the dictionary of collocation rules, on the basis of the previous studies only considering the synonyms of opinion words, this paper proposes to introduce the antonyms of opinion words to expand the rule dictionary. By comparing with the traditional methods, the algorithm proposed in this paper is proved to be true and effective, and is a good result in the field of implicit feature extraction.

2. An implicit feature recognition method based on antonym expansion of collocation rule dictionary

2.1 Preprocessing

The preprocessing of text includes clauses, that is, long comment sentences are divided into clauses by commas or periods according to the separation of sentence meaning in Chinese, because each clause in Chinese comment means an independent meaning. Then there is the word segmentation and part-of-speech tagging, which this article USES as the Jieba word segmentation, and all data manipulation and computation is done in Python code. After preliminary data preprocessing, features and opinion words need to be defined. The so-called features are the actual objects described by this clause. For example, in the hotel field, the features frequently reviewed in the user review clause include price, environment, service, etc., while in the mobile phone field, the features frequently reviewed by users include appearance, price, screen, battery, etc. The meaning of an opinion word is the emotional tendency and attitude of a certain feature. For example, in the comment "This price is good and affordable", "price" is the feature and "affordable" is the opinion word. In the study of this paper, the most general case is considered, namely, the noun is set as a feature word and the adjective as an opinion word. Meanwhile, according to Zhang Li in his paper, words such as "good" and "good" can often be matched with a variety of features, so there is no discriminability and one-to-one mapping relationship between opinion words and features. Therefore, this paper does not directly consider such opinion words that can point to a variety of features.

In traditional methods, comments need to be labeled into explicit and implicit comments before experiment, so this paper proposes a rule to filter explicit and implicit

comments, which can effectively reduce the time of many manual annotations [错误! 未定义书签。]:

1) Determine if an adjective exists in the comment clause, and delete the clause if it does not exist or if 3 or more adjectives exist. (Lack of opinion words or too many)

2) Determine if there is a noun in the comment clause. If there is no noun, it is an implicit comment. If there are 3 or more, the clause is deleted. In other cases, the clause is a display comment. The main processes of preprocessing are shown in Figure 1 below:

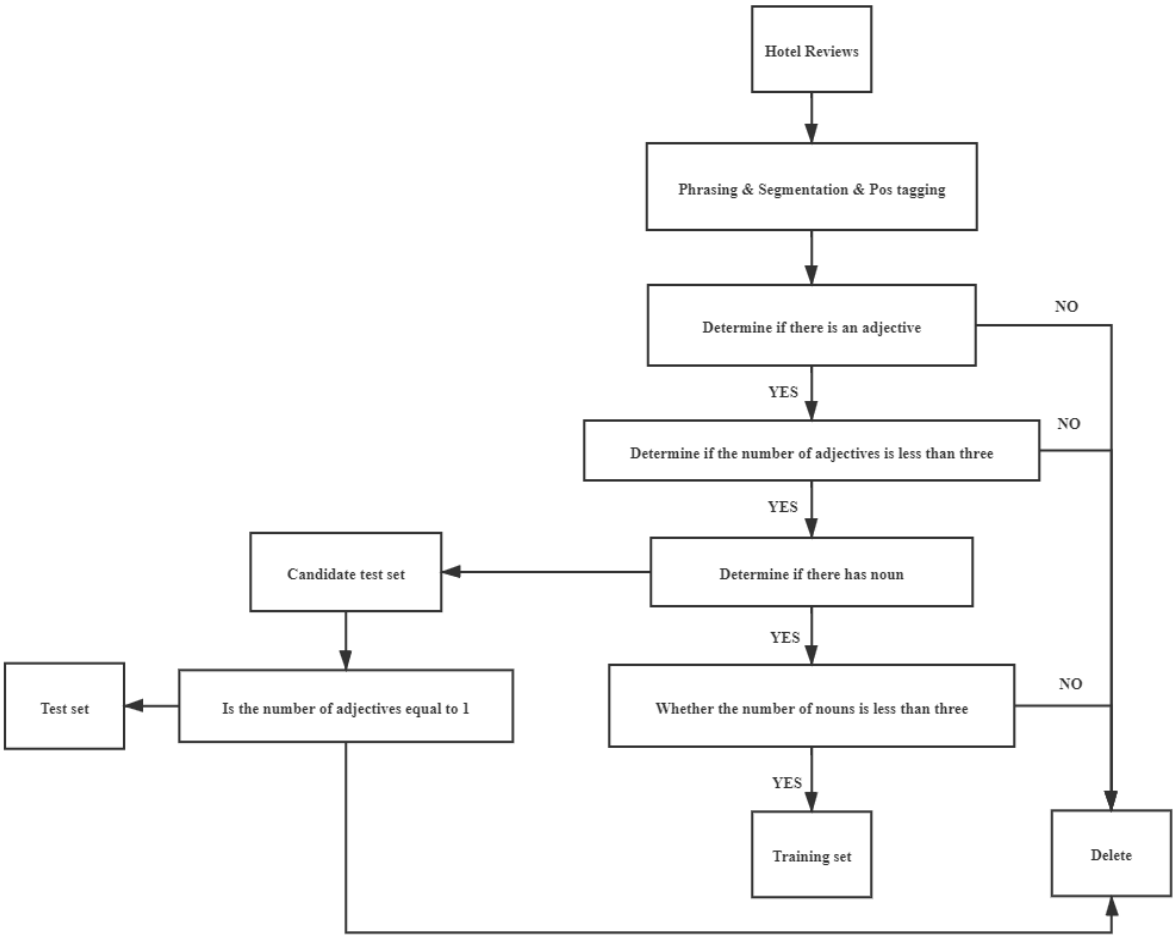

Figure 1 Preprocessing main flow

2.2 Synonym merging of feature words

In voluminous comments, due to the characters of the extensive and profound, also often have a lot of meaning of the word synonymous expressions such as "room" and "house", in order to avoid repeat redundant rules dictionaries match rules, so it needs to be synonymous with key words are merged into and then choose one of the most representative of the other synonyms replacement, however, how to identify the semantic difference itself is a very challenging thing, but it can from the view of the similarity of feature to measure the possibility of synonyms, the commonly used method is as follows:

1) Think of features as strings. Since the nouns in Chinese are composed of Chinese characters, it can be judged from this that if the two words have Shared character strings, they are likely to be synonyms. For example, "room", "house" and "apartment" all means "room", so they are classified as synonyms and replaced by the most representative "room".

2) Two features in the comment of the co-occurrence frequency can also be measured for the basis of synonyms, because according to the prior knowledge can be analyzed out, in Chinese, usually do not at the same time with two different description to express the same characteristics, so the co-occurrence of key similarity is smaller, the more the more can't be synonyms, conversely, the greater the chance of synonyms.

Based on the above two methods, a formula can be used to calculate the similarity between the two features, as shown in (1):

$$Sim(f_i, f_j) = \frac{length(substring(f_i, f_j))}{\max(co-occurrence(f_i, f_j), 0.5)} \quad (1)$$

Where, the numerator is the length of the Shared substring of the two features, the denominator is the number of co-occurrences if the two features appear in the same comment, and the value is 0.5 if the two features have not co-occurred. A threshold MinSim is then set. If the similarity between the two features is greater than the threshold, it will be a synonym.

2.3 Construct binary groups

Because the comment clause is composed of many words, among which the most critical information is opinion words and feature words, the dualistic group of construction (feature, opinion words) is used to represent the information of the sentence [12], The collocation rule of binary group is to match the feature words and opinion words in the clause in turn.

After constructing the set of binary groups, the binary groups with occurrence frequency of binary groups greater than 3 times are screened out. Because the binary groups with co-occurrence frequency lower than 3 times bring great noise, the screened binary groups are used to further construct the collocation dictionary.

2.4 Creating a dictionary

In line with most mainstream approaches, this paper use display comments with explicit features to construct pairs of collocation rules in the form of several (feature-opinion) binary groups [13], Then the implicit feature is identified by the opinion words in the implicit comment sentence and the mapping relationship in the dictionary of collocation rules.

However, in this article, there are some differences compared to the existing methods in building the dictionary:

1) According to the second-level collocation rule table formed after the clustering of

feature words, while the mainstream methods only consider the synonyms of opinion words[14] to expand the collocation rule pair, consider the antonyms of opinion words to add the collocation rule pair, and then write these collocation rule pairs into the dictionary for storage. For example, in the hotel domain review, "clean" and "dirty" are opposites. If the only collocation rule in the dictionary is (room-clean), add (room-dirty) to the collocation rule pair.

2) The mainstream method does not involve the operation of removing noise opinions. Found in the process of practical experiments, for example: "good", "well" and "big", "small" and "fit", can instruct many types of completely different key words, will bring the results a lot of noise, cannot correct mark in the test set of annotation of key word so to consider the noise removal.

3) After the step of appeal, according to the opinion words in the test set, the word with the highest PMI value is found as its pointed feature word, and then stored in the dictionary of collocation rules. PMI [14] is a formula to measure the interdependence of two words. The calculation formula of PMI is as follows:

$$PMI(o_1,\ f_1) = \log \frac{p(o_1 f_1)}{p(o_1)p(f_1)} \tag{2}$$

Where, o_1 is an opinion, f_1 is a feature word, $p(o_1)$ is the probability of an opinion word, $p(f_1)$ is the probability of a feature word, $p(o_1f_1)$ is the co-occurrence probability.

2.5 Implicit feature recognition

For the test set, an implicit comment that meets the conditions mentioned in the preprocessing section is obtained after processing, and then corresponding features are identified through the opinion words in the sentence and the mapping relationship in the dictionary of collocation rules, and the feature words obtained are judged as the implicit features of the sentence.

3. Experimental results and analysis

3.1 Experimental data acquisition and preprocessing

The experimental data of this paper is Tan Songbo's hotel comment corpus, among which 6,267 comments are randomly selected, and 75,759 clauses follow clauses. After word segmentation and part-of-speech tagging, 7755 explicit comment clauses and 4,245 are selected as the test set after screening through preprocessing section.

3.2 Experimentation

1) Synonym merging of feature words

In the training set, the synonyms of the feature words are merged first, as shown in Figure 2, and the synonyms are replaced with the most general words:

Bed	Bed
Sleep bunk	

Figure 2 Example of a synonym merge

2) Generate a dictionary of collocation rules

By calculating the sentences in the training set line by line, 2344 binary collocation rule groups (features, opinion words) with co-occurrence times greater than or equal to four times are obtained, and 25 candidate implicit features are obtained, as shown in Figure 3:

Candidate implicit features											
Time	Room	Serve	Price	Hotel	Reception	Traffic	Feel	Breakfast	City	Facility	Location
Environment	Quilt	Bed	Information	Speed	Toilet	Waiter	Restaurant	Quantity	Light	Water	

Figure 3 Candidate implicit features

3) Implicit feature recognition

The collocation rule dictionary with the shape of {opinion words, features} is constructed by calculating the feature words with the most co-occurrence times of opinion words, and then synonyms and antonyms of opinion words are introduced to expand the collocation rule dictionary. Examples are shown in Figure 4 and Figure 5 below:

Expensive	Price

Figure 4 Example of the original collocation rule

Expensive	Price
Costly	
Exorbitant	
Cheap	
Inexpensive	

Figure 5 introduces synonyms and antonyms to expand collocation rule pairs for examples

3.3 Experimental results of implicit feature recognition

Opinion words were extracted from 4245 comment data, and the implicit feature words were determined according to the mapping relationship between opinion words and feature words in the dictionary of collocation rules. The experimental results of this paper were compared with the method in literature 11[15], and the final results were shown in Figure 6 below:

Precision	Recall	Method
0.876	0.866	This text
0.854	0.751	References 11

Figure 6 Experimental results

Precision

$$\text{precision} = \frac{C_{\text{r}}}{C_r + C_w} \tag{3}$$

Where, C_r is the number of consistent rows corresponding to machine recognized columns and manual labeled columns, C_{w} is the number of inconsistent rows corresponding to machine recognized columns and manual labeled columns (blank rows are not involved in calculation).

Recall

$$recall = \frac{C_{\text{z}}}{C_{\text{non}} + C_{\text{z}}} \tag{4}$$

Where, C_z is the number of results found by the machine recognition column, and C_{non} is the number of empty rows in the machine recognition column.

4. Conclusion

It can be seen from the experimental results that the algorithm proposed in this paper is good, with high accuracy and recall rate. Especially after the introduction of antonyms, the recall rate has been improved by 11.5%. Moreover, the algorithm proposed in this paper is relatively simple to implement, with low calculation cost and less manual intervention, which is of high practical value.

However, this paper still has limitations. The research done in this paper is aimed at the field of hotel review, and the implicit feature recognition in the general field needs further research.

References

[1] Qi Su, Xinying Xu, Honglei Guo. Hidden sentiment association in chinese web opinion mining[C]// World Wide Web Conference. 2008.

[2] GAWADE J, PAETHIBAN L. Opinion mining feature extraction using domain relevance[C]//Proceedings of the International Conference on Data Engineering and Communication Technology. 2017:401-409

[3] PORIA S, CAMBRIA E, GELBUKH A. ASPECT extraction for opinion mining with a deep convolutional neural network[J]. Knowledge-Based System, 2016,108:42-49.

[4] Li Shi, Ye Qiang, Li Yijun, Rob Law. Research on product Feature Mining method of Chinese Online customer Comments [J]. Journal of Management Science, 2009, 12(02): 142-152.

[5] Li Changbin, Pang Chongpeng, Ling Yongliang, Wang Qiang. Research on Network Comment Mining based on Improved Feature Extraction and Clustering [J]. Modern Intelligence,2018,38(02):68-74.

[6] Li Changbin, Pang Chongpeng, Li Meiping. Application of Apriori Algorithm based on Weight in Text Statistical Feature Extraction method [J]. Data analysis and knowledge discovery,2017,1(09):83-89.

[7] Wang Fudong, Yin Qianqian, Liu Fengtao. Implicit Recognition of Commodity Characteristics in online reviews [J]. Journal of Donghua University (Natural Science edition),2019,45(03):451-456.

[8] LIU B, HU M, CHENG J. Opinion observer: Analyzing and comparing opinions on the web[C]//Proceedings of the 14th international conference on World Wide Web. ACM, 205:342-351

[9] XU H, ZHANG F, WANG W. Implicit feature identification in Chinese reviews using explicit topic mining model[J]. Knowledge-Based Systems, 2015,76:166-175.

[10] Zhang Li，Xu xin. Research on implicit Attribute Extraction in Product Reviews [J]. Modern book Information Technology, 2015,31(12):42-47.

[11] Liu Hongyan. Social Computing: Analysis and Mining of Users' online Behavior [M]. Beijing: Peking University Press

[12] Qiu Yunfei, Ni Xuefeng, Gao Niangshan. Research on the Method of extracting commodity implicit Evaluation object [J]. Computer Engineering and Application,2015,51(19):114-118.

[13] Jiang Tengjiao, Wan Changxuan, Liu Dexi, Liu Xiping, Liao Guoqiong. Evaluation object - Emotion word pair extraction based on semantic analysis [J]. Acta Computerica Sinica,2017,40(03):617-633.

[14] Tian Jiule, Zhao xin. A Method for word Similarity Calculation based on Thesaurus Forest [J]. Journal of Jilin University (Information Science edition),2010,28(06):602-608.

[15] Yang xin, Yang Yunfan, Jiao Wei, Zhu Donglin, Zheng Shaoyang, Yuan Zhongyu, Yang Xiuzhang, Luo Zijiang. Affective analysis of home stay Review based on domain Dictionary [J]. Science, Technology and Engineering,2020,20(07):2794-2800.