Sentiment Mining Analysis Based on Network Comments

Chengfang Tan[1, 2]

[1] *School of Information Engineering, Suzhou 234000, Anhui, China*
[2] *Intelligent Information Processing Lab, Suzhou 234000, Anhui, China*

Guolong Chen, Qixiang Song

School of Information Engineering, Suzhou 234000, Anhui, China

Abstract
In order to solve the problem of lacking of semantic understanding and insufficient sentiment analysis in traditional sentiment analysis method, this paper analyzes the sentiment tendency of network comments text from the semantic point of view. Firstly, based on the existing sentiment dictionary, we sort and build a sentiment dictionary for network comments, and analyze the emotional effect of degree adverbs, negative words and punctuation on the comment text. Then by extracting co-word and calculating semantic co-word matrix, we analyze the different product features sentiment tendency from the dimension of user concerned. Finally we obtain the sentiment tendency of each comment text through weighted calculation. Experimental results show that the proposed sentiment mining method for network comments has obvious analysis effect.
Key words: NETWORK COMMENTS; SENTIMENT DICTIONARY; SENTIMENT MINING; SENTIMENT TENDENCY

1. Introduction

With the rapid development of the Internet, forums, microblog and other communication platform are increasingly emerging. People are accustomed to publication of the network comment on buying goods. Those comments have short length and evident emotions. Based on the sentiment analysis of network comments, businesses can understand the shortage and advantage of themselves and competitors to make the right decisions. At the same time it also can capture consumer preferences from these comments, and provide data support for the direction of future development. Sentiment mining, also known as opinion mining, is to find the commodity's attitudes and opinions of the consumers on product by automatically analyzing the product comments text.

Pang et al. used machine learning methods for tendency classification of movie reviews on Usenet [1]. Chenxi X improved the effect of opinion mining with the establishment of corpus and the tree of the knowledge [2]. Turney P used mutual information to determine the sentiment tendency of users, got rid of dependence on sentiment dictionary, and proved the effectiveness of the method through on car reviews data source experiment [3]. At present, there are many researches on sentiment analysis of network comments, but there are also some problems. For its short text, traditional text

classification algorithm is limited its role. Sentiment mining limited to analysis of overall satisfaction with products, not from the users concerned dimensions. As the network comments are unstructured text information and has domain knowledge, therefore, how to carry on the thorough excavation to find sentiment tendency and its strength, text mining puts forward new challenges。

This paper researches sentiment mining for network comment text. It builds the sentiment dictionary for comment text, fully considers the effect of degree adverbs, negative words and symbols on emotional words, and extracts co-word and calculates semantic co-word matrix to achieve the sentiment tendency analysis on product comments.

2. Construction of the sentiment dictionary

According to the characteristics of network comment text, we construct the sentiment dictionary which includes basic dictionary, field dictionary, network dictionary, and modified words dictionary.

(1) Basic dictionary.

This paper uses the emotional words set released by HowNet as the basis [4], according to the provided emotional words from “National Taiwan University sentiment dictionary" and "Students appraise meaning dictionary” [5], organizes them as basic dictionary after the deduplication.

(2) Field dictionary

Some polarity words are only used in specific areas, and have sentiment tendency, such as "limit", and some polarity words will show different sentiment when modifying different characteristics in different areas.

(3) Network dictionary.

Large numbers of network terms are often used to express people’s sentiment tendency over a period of time, such as "Rice porridge”, and so on. Therefore, in order to meet the needs of network comment, those network terms with sentiment tendency used frequently are added to sentiment dictionary.

(4) Modified words dictionary.

The length of network comment text is short and informative, non-documented style of writing often uses the modifier to modify emotional expression of the user. When the degree adverbs or negative adverbs modify emotional words, the sentiment polarity and strength are likely to change.

3. Clustering analysis of network comment text

This paper introduces domain ontology, fully considering the semantic relation between two words, calculates the semantic similarity of co-word. Word text set $D=\{d_1,d_2,...,d_m\}$, d_i represents the *i*th short texts, and f_i represents the weight of the *i*th short texts. Word frequency set $F=\{f_1,f_2,...,f_s\}$, w_i represents the word frequency of the *i*th word in text set D.

3.1. Extraction of high frequency words

The text set D is processed with word segmentation, calculates the word frequency f_i of each word, removes stop words and useless words, and sets threshold U_1, delete the word w_i when $w_i < U_1$.

3.2. Construction of co-word matrix

Sequentially calculate the co-occurrence frequency of high frequency words, denoted as f_{ij}. The co-occurrence frequency of words multiplies with the weight value w_i that is the support number of each comment. In order to obtain better clustering effect, we use the co-occurrence strength to replace the absolute co-occurrence frequency [6], which is calculated as follows.

$$s_{ij}=\frac{f_{ij}\times w_i}{\sqrt{f_i\times f_j}} \quad (1)$$

3.3. Construction of semantic co-word matrix

With the introduction of domain ontology, semantic similarity calculation of co-word can be calculated. This paper uses the concept similarity calculation method. The formula is as follows [7].

$$\sin(i,j)=\frac{1}{m}\sum_{n=1}^{m}\delta_n(i,j) \quad (2)$$

Where i and j respectively represent two words of a set of words, m is the larger path value of i and j in the domain ontology. It needs to traverse the parent class concept of the front 1, 2... N of i and j, if the parent class concept is the same, $\delta_n(i,j)$ is counted as 1, the other is counted as 0.

The calculation formula of the semantic co-occurrence frequency is as follows:

$$s_{ij}=\frac{f_{ij}\times w_i}{\sqrt{f_i\times f_j}}+\frac{1}{m}\sum_{n=1}^{m}\delta_n(i,j) \quad (3)$$

Set the threshold U_2, delete the words where $s_{ij} < U_2$, normalize the results, and construct the semantic co-word matrix, which is as follows.

$$\begin{bmatrix} S_{1,1} & S_{1,1} & S_{1,1} & \cdots & S_{1,i} \\ S_{2,1} & S_{2,2} & S_{2,3} & \cdots & S_{2,i} \\ \vdots & \vdots & \vdots & & \vdots \\ S_{j,1} & S_{j,2} & S_{j,3} & \cdots & S_{j,i} \end{bmatrix}$$

3.4. Cluster analysis

The semantic co-word matrix is imported into data analysis software to perform quantitative

cluster analysis. Through cluster analysis, we can get the main features that the user comments concern on, statistics the comment texts for each feature, and determine all keywords of the feature and construct classification vocabulary. Traditional clustering analysis tools have SPSS, SAS and so on.

4. Sentiment mining process of comment text

4.1. Sentence division

Each comment text usually contains the evaluation of two or more than two kinds of different features for comment product. Therefore, in order to accurately reflect the specific evaluation on each feature of the product, it is necessary to analyze the sentiment tendency according to the different features of product. This paper selects the punctuation in Chinese sentence"." ,"!" and "..." as the marker of the end of sentence, taking into account the normative symbols used in the network text, also adds other symbols as auxiliary marker, such as "," and so on. Use formal style to build matcher sentence segmentation to scan each comment text in turn and store text as the unit of sentence.

4.2. Sentiment tendency analysis of degree adverbs, negative words and punctuation

Degree adverbs, negative words and punctuation play different effects on tendency of sentence, which can strengthen or weaken the sentence emotion. In order to accurately analyze the emotional intensity of user comments, this paper sets up a detection window in the context of emotional words, where the detection window size is 5. If the degree adverb w_a modifies emotional word w_i in detection window, the sentiment tendency calculation method of emotional word w_i is as follows.

$$o(w_i) = M_{w_a} * S_{w_i} \quad (4)$$

Where M_{w_a} represents the strengthen of degree adverbs, S_{w_i} represents the weight of emotional word w_i.

If the negative word modifies emotional word w_i in detection window, the sentiment tendency calculation method of emotional word w_i is as follows.

$$o(w_i) = (-1)^n * S_{w_i} \quad (5)$$

Where n represents the number of negative words modified emotional word w_i in detection window, S_{w_i} represents the weight of emotional word w_i.

In this paper, emotional strength of sigh is set to 2, and the question mark is set to -2. When the sigh or question mark modifies the emotional words w_i, the sentiment tendency calculation method of emotional word w_i is as follows.

$$o(w_i) = M_{pun} * S_{w_i} \quad (6)$$

Where M_{pun} represents the strengthen of the sigh or question mark, S_{w_i} represents the weight of the emotional word w_i.

4.3. Sentiment tendency analysis of sentence

When performing the calculation of sentiment tendency, this paper is based on sentence segmentation and POS tagging to judge whether there are emotional words. If there is emotional words, gets its emotional intensity, then detects whether there are degree adverbs and negative words appear in detection window, if there is, according to the formula (1), formula (2) to treat. Punctuation at the end of the sentence is treated in accordance with the formula (3). Therefore, sentiment tendency calculation method of sentences s_i is as follows.

$$o(s_i) = \sum_{i=1}^{k} o(w_i) \quad (7)$$

Where k represents the number of emotional words contained in sentence.

4.4 Sentiment tendency analysis of comment text

Assuming the comment text d_i contains n sentences, so the sentiment tendency calculation method of the comment text is as follows.

$$o(d_i) = \sum_{i=1}^{n} o(s_i) \quad (8)$$

According to the formula (5), we can get the sentiment tendency value $o(d_i)$.it will be one of the following three conditions: $o(d_i) > 0$, $o(d_i) = 0$, $o(d_i) < 0$. Therefore, according to the different situation of the value $o(d_i)$, we can identify the sentiment embodied in the comment text is positive, negative or neutral.

5. Experiments and Analysis

5.1. Experimental Data

To verify the proposed sentiment tendency calculation method in this paper, we use the open-source crawlers Soukey to get several brand mobile phone comment content from Zhongguancun online. A total of 800 comments text, half of which is used for training, half for the experiment. We mark the product features contained in each comment text by artificial

means and give the positive, negative, or neutral emotion annotation, and also judge the sentiment tendency on each comment text.

5.2. Evaluation Index

In this paper, we use information retrieval evaluation indexes which are widely used for evaluating the experimental results. Evaluation indexes are Precision (P), Recall (R), and F-measure [8]. The formula is as follow.

$$P = \frac{a}{a+b} \quad (9)$$

$$R = \frac{a}{a+c} \quad (10)$$

$$F = \frac{2RP}{R+P} \quad (11)$$

Where a represents the number of texts which are assigned to a category correctly, b represents the number of misclassified texts in the classification results, c represents the number of texts that should be assigned to a category, but there is no proper classification, k represents the number of categories of texts.

5.3. Analysis of Experimental Results

This paper uses Institute of Computing Technology of Chinese lexical analysis system (ICTCLAS) to achieve Chinese word segmentation and speech tagging, filters out all the nouns and adjectives, statistics word frequency and obtains high frequency words, and then extracts the co-word pairs and calculates the frequency. The semantic co-word matrix can be obtained with the introduction of domain ontology. Three words are selected as the main features of the product comments through cluster analysis. Determine all the keywords of the feature and construct classification word table, the experimental result is as shown in Table 1.

Table 1. Keywords of three features

Feature	Corresponding Keywords
Style	appearance, screen, body, color, size, quality, thin.....
Function	quality, capacity, reaction, call, pixels, signals, battery .....
Service	return, warranty, speed, attitude, express delivery, customer service, packaging .....

After extracting features of comment texts and classifying sentences, we analyze sentiment tendency of different features to obtain user's attitude for the product. The experimental result is as shown in Figure1.

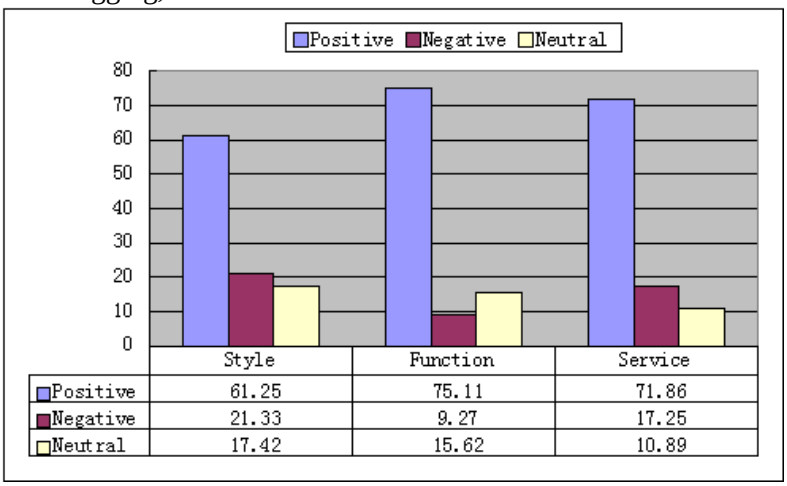

Figure 1. Sentiment tendency result of different features

From Figure 1, we can see clearly that users give total evaluation for the product, including design, function and service of the main 3 characteristics. The evaluation result provides an important reference for other users.

This paper selects 600 product comment texts, which contains 386 positive comments, 91 negative comments, and 123 neutral comments. This experiment use 3 kinds of classification methods for comparison, namely support vector machine (SVM), simple Bias (NB) and the proposed method based on sentiment dictionary (SD). Among them, SVM uses a standard tool light-SVM, and NB uses the Mallet machine learning toolkit. When using these tools, all parameters are set to their default values. Comparative experimental results are as shown in Table 2.

From the comparison experiment, it shows that three different experiment methods have obvious difference. Based on the sentiment dictionary, the precision rate, recall rate and F-measure are significantly improved after introducing the analysis of negative words, degree

adverbs and punctuation. The proposed method in this paper is proved to obtain a better sentiment tendency analysis results on comment texts.

Table 2. Comparison results of three different methods

Experimental method	Precision (%)	Recall (%)	F-measure (%)
SVM	78.85	79.23	79.04
NB	81.29	80.57	80.93
SD	84.67	85.91	85.29

5. Conclusions

This paper analyzes the sentiment tendency of network comment texts. Based on the sentiment dictionary constructed, we extract the main features of product by semantic co-word matrix and its clustering analysis. Finally, the sentiment tendency mining for different features and all comment text are achieved respectively. Experimental results show that the method proposed in this paper has higher accuracy in the sentiment tendency analysis, which provides a convenient and practical guidance for users, manufacturers and decision-making. As the words and structure of network comment texts are not standardized, and Chinese expression is more subtle, which have a certain impact on the analysis results, so the calculation rules need to be improved in the next work.

Acknowledgments

This work was supported by Key University Science Research Project of Anhui Province (No.KJ2014A247) and Major University Science Research Project of Anhui Province (No.KJ2014ZD31) and Open Project of Intelligent Information Processing Lab at Suzhou University of China (No.2014YKF41)

References

1. Pang Bo. and Lee, Lillian. (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proc. Conf.on the Association for Computational Linguistics (ACL)*, volume 43, p.p.115-124.
2. Chenxi X. (2012) Competitive intelligence system based on data mining view. *Journal of intelligence*, 31(2), p.p.174-179.
3. Turney P. (2002) Semantic orientation applied to unsupervised classification of reviews. *Proc. Conf. on the 40th Annual Meeting of the Association for Computational Linguistics*. Momstown, HJ, USA, p.p.417-424.
4. Zhu Y, Min J, Zhou Y, et al. (2006) Semantic orientation computing based on HowNet. *Journal of Chinese Information Processing*，20(1), p.p.14-20.
5. Dang, L., Zhang, L. (2010) Method of discriminant for Chinese sentence sentiment orientation based on HowNet. *Application Research of Computers*, 27(4), p.p.1370 - 1372.
6. CUI H, MITTAL V，DATAR M. (2006) Comparative experiments on sentiment Classification for online product reviews. *Proc. Conf. on the 21st National Conference on Artificial Intelligence*,p.p. 1265-1270.
7. Matsunoto S,Takamus,Okumur (2005) Sentiment classification using word sub-sequences and dependency sub-trees. *Proc. Conf. on the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Berlin: Springer, p.p.301-310.
8. Chengfang Tan (2013) Sentiment Tendency Analysis of MicroBlog Based on Semantic. *International Journal of Applied Mathematics and Statistics*, 51(21), p.p.501-509.