The Sentiment Analysis in Opinion Mining of Web News Comments

Qing Du[1, a], Fen Guo[2, b] ,Jiangxing Wang[3, c]

[1] School of Software Engineering, South China University of Technology, Guangzhou, 510006, China

[2] School of Software Engineering, South China University of Technology, Guangzhou, 510006, China

[3] School of Software Engineering, South China University of Technology, Guangzhou, 510006, China

[a]duqing@scut.edu.cn, [b]csguofen@scut.edu.cn, [c]598547082@qq.com

Keywords: Opinion Analysis; Semantic Analysis; Semantic Similarity Calculation

Abstract. Opinions' analysis on web news comments is a great significance for the construction of the government's democratic politics, public opinion monitoring and decision support, etc. In this paper we analyze user's opinion from emotional word, emotional word qualifier, and emotional tendency. We proposed an emotional word qualifier extract algorithm based on dependency relation tree. We improved the words polarity calculation algorithm based on HowNet. Finally, we implemented the system of opinion mining on web news comments, and got the overall sentiment analysis.

Introduction

The research objects of sentiment analysis in this paper are web news comments. The comments of the internet news are an important feedback channel of the public political supervision. So how to extract the views of the public from massive web news comments quickly and effectively? The opinion mining of the web news comments is becoming an important part of the public opinion monitoring. And we describe the opinion of a comment using quaternion which includes characteristic, emotional words; embellish words of emotional words and emotional tendencies.

In the work of sentiment analysis, BoPang[1] classified the movie evaluating data of IMDB by using 3 machine learning technologies (Naive Bayesian Model, The maximum entropy, Support Vector Machine), and got the polarity classifier; Hu and Liu[2] proposed the polarity recognition when the emotion word is an adjective; Turney[3] observed that words with the same polarity usually appear in one sentence, so it can be used that using PMI-IR to recognize the polarity of words; Xiao[4] classified blog paper by using Naive Bayesian Model Support Vector Machine and Rocchio Classification Algorithm; Xu[5] recognized the sentiment polarity of an emotion word by calculating the similarity of unknown emotion word and defined polarity word seeds using HowNet.

We proposed the algorithm based on chain dependency relationship to extract the qualifier, improving the accuracy of word sentiment polarity recognition based on HowNet by optimizing the sentiment word dictionary. Also we quantify the polarity analysis, so that we can classify the sentiment of the opinion in news.

The Extraction of Emotional Word's Qualifier

Lin[6] pointed out that the degree adverb has the following three characteristics:(1)Strong adhesion;(2)Great positioning capability, can only be used as adverbial;(3)Single semantic direction , always point to the qualified word.

Summing up those characteristics, we can find that in the syntax dependency tree:

（1）The qualifier is adverbial;

（2）There is direct or indirect dependency relationship between the qualifier and the emotional qualified word. For example, in the comment "This policy is not correct", the emotional word is "correct", "not" is the qualifier, and the relationship between them is the adverbial relationship. If there are multiple qualifiers, the relationship between the qualifier and the emotional qualified word will be the chain dependency relationship. For example, in the comment "This policy is extremely not correct", their syntax dependency tree is shown in Fig 1. The negative word "not" and the emotional word "correct" have the dependency relationship, and degree words "extremely" and "not" also have the dependency relationship. They form the chain dependency relationship, as shown in Fig 1 a). Furthermore, this chain relationship also exist in the comment "This policy is very very not correct", which is shown in Fig 1 b). So we firstly find all the words which is adverbial along the dependency chain, then we can find all the dependency word of the emotional word.
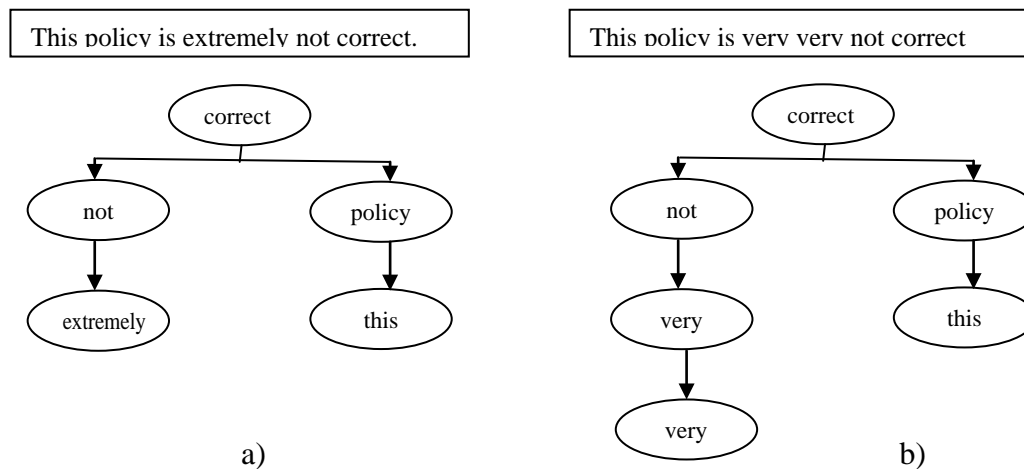

Fig1 The Example of Syntax Dependency Tree and Chain Relationship

Improved HowNet Sentiment Polarity Recognition Algorithm

The sentiment polarity recognition of words is determined by the semantic similarity between the word and the benchmark word. We use the method of semantic similarity calculation based dictionary. In Chinese, the Hownet is used most. To improve the recognition rate of HowNet algorithm , we build up a new emotional word dictionary as the set of benchmark words, the steps are as follows：

（1）Firstly , get 4360 commendatory terms and 4565 derogatory terms from HowNet;

（2）Then , check these words by using HowNet , eliminate the entry which is not in the HowNet. If the word is not in the HowNet, it will affect the calculation accuracy. After the elimination, there are 3254 commendatory terms and 2939 derogatory terms, totally 6193 words;

（3）Finally , we make a synonym judgment on all those 6193 words , and hold only one word. After this step, there are 740 commendatory terms and 734 derogatory terms, totally 1474 words.

Label the polarity of those 1474 words manually, given commendatory terms the polarity value -1 and derogatory terms 1.

Based on the works above-mentioned，we purposed the algorithm to recognize the emotional word's sentiment polarity in this paper, the algorithm steps is shown as followed:

Algorithm: The Emotional Word's Sentiment polarity Recognition

Input：Emotional word

Output：Relevant polarity

Ensure：

（1）For the emotional word (OpinionWord) , use the correct list to find and correct the ghost words , turn them into the correct emotional word , OpinionWord -> CorrectedOpinionWord；

（2）For the CorrectedOpinionWord , do the synonym expansion , get the word list (WordList$\{Word_1, Word_2 \ldots Word_n\}$) which also contains the emotional word itself.

（3）For the WordList$\{Word_1, Word_2 \ldots Word_n\}$, and the emotional word dictionary (SenWordDict$\{$ $SenWord_1.SenWord_2 \ldots SenWord_m\}$) , use the method of semantic similarity calculation based on HowNet , calculate the semantic similarity between words in the Wordlist and the emotional word dictionary, $Sim_{i=1,..n; j=1,...m}(Word_i, SenWord_j)$ ；

（4）For the biggest semantic similarity we get $Max(Sim_{i=1,..n; j=1,...m}(Word_i, SenWord_j))$, if it's bigger than a certain threshold $\theta_{threhold}$,we regard the polarity of emotional word (SenWord) where we get the biggest similarity as the polarity of emotional word (OpinionWord). Polarity （OpinionWord）=Polarity（SenWord）. If the word is not emotional word, it has no sentiment polarity.

Sentimental Analysis of the Entire Comment

In order to get the polarity of the entire comment, we give a strength value to every word in the degree word dictionary and negative word dictionary. As for adverbs of degree, strength value reflects the adverbs' additional effects of sentiment words. If the words enhance the sentiment, the value is positive. For example, the word "very" has the value of 0.5. On the contrary, for words used to weaken the sentiment, the value is negative. Like the word "slightly" , its value is -0.2. Moreover, for privatives, which represent the opposite effect towards polarity of sentiment words, their value is always -0.8.

Here is the steps to calculate the polarity of the quaternion <characteristic, emotional word, embellish words of emotional words, emotional tendencies >：

1) Give sentiment word's polarity to the quaternion as an initial value, $P_{tuple} = P_{sentiment\_word}$ ；

2) According to the modifier list of the sentiment word, do following judgments from the closest qualifier:

a. If it is an adverb of degree, using this formula to adjust the value:

$$P_{tuple} = \begin{cases} -s_{modify\_word} + P_{tuple}, P_{tuple} < 0; \\ s_{modify\_word} + P_{tuple}, P_{tuple} > 0; \end{cases} \tag{1}$$

b. If the modifier belongs to privative, using this formula to adjust the value:

$$P_{tuple} = \begin{cases} s_{negate} \times (P_{tuple} - 1), P_{tuple} > 1; \\ s_{negate} \times P_{tuple}, P_{tuple} \leq 1; \end{cases} \tag{2}$$

Summing up, the method used in this paper to calculate the polarity is like the following Table 1.

Table 1: Example of Comment Polarity Calculation

Example	The calculation process
I like this policy	(1) $P_{\text{tuple}} = P_{sentiment\_word} = 1$
I don't like this policy	(1) $P_{\text{tuple}} = P_{sentiment_{word}} = 1$ (2) $P_{\text{tuple}} = s_{negate} \times P_{\text{tuple}} = -0.8 \times 1 = -0.8$

For the result of the polarity calculation , we divided the different value scope into five classifications according to Table 2 : Very negative、Negative、Neutral、Positive、Very positive.

Table 2: Natural sentence polarity classification

Strength	Polarity range
Very positive	(1.2,)
Positive	(0.2,1.2]
Neutral	[-0.2,0.2]
Negative	[-1.2,-0.2)
Very negative	(,-1.2)

Test Results

Use the improved algorithm to do the experiment, the result is shown in Table 3.

Table 3: The experimental result of the algorithm to judge the word sentiment polarity in this paper

Classification	HowNet			Improved HowNet		
	Prec%	Rec%	F %	Prec%	Rec%	F %
commendatory	80.7	78.5	79.6	86.3	83.8	84.5
derogatory	81.5	78.0	79.7	87.7	82.8	85.1
overall	81.1	78.3	79.6	87.0	83.2	84.8

We can see that after the benchmark set extension and the synonymous extension, the algorithm's recall has improved about 5%, and the using of threshold filtering also leads to an improvement of algorithm's precision, which has up to 86%.

Conclusion

Opinion mining about web news comments has an important effect on government's public opinion monitoring and news agency's public opinion analysis. In this paper, we improve the precision and recall of the characteristic-emotional word pair extraction by expanding the search range and filtering the characteristic set. In the future, consider storing the comment data in the data warehouse to satisfy the storage and analysis of mass data, we should use distributed computing platform like Hadoop to get a better performance on the management of the comment data.

References

[1]Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques[C]. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. 2002:79-86.

[2]Hu M, Liu B. Mining and summarizing customer reviews[C]. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM

Press,2004: 168-177

[3]Turney PD: Mining the Web for Synonyms PMI-IR Versus LSA on TOEFL. Berlin:Springer-Verlag,2001

[4]Xiaochuan Ni, Gui-Rong Xue, Xiao Ling: Exploring in the weblog space by detecting informative and affective articles[J]. Track: Industrial Practice and Experience,2007(8-12):281-290

[5]XU L, LIN H, YANG Z:Text orientation identification based on semantic comprehension[J]. Journal of Chinese Information Processing, 2007, 1: 015.

[6]Linheng: , The Scope and Characteristics of Adverbs of Degree Classification, Journal of Shanxi University (Philosophy and Social Sciences), 2006(26).71-74