A Content-Based Blogger's Sentiment Analysis System

Guohua Ou [1, a *], Xiaoyan Zhuo [1,b] and Wu Tang [1,c]

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

[a] csghou@scut.edu.cn, [b]563156382@qq.com, [c]710102161@qq.com

Keywords: Sentiment analysis; Comment target extraction; Term-frequency filtration; NN filtering algorithm

Abstract. Nowadays, the information of personal viewpoints booms on the Internet. The sentiment analysis based on the content of blogs is a hot issue in natural language processing. This paper establishes a system of blogger's sentiment analysis which is based on comment target extraction. It is supposed to extract the sentimental information and key events in the blogs as well as the related characters, time and locales. This paper makes an intensive study of its key technology – comment target extraction, and presents a target filtering method based on term-frequency filtration and NN filtering algorithm. It also classifies sentences of sentiment and makes analysis rules for that. According to test results, the system has higher accuracy.

Introduction

With the development of the Internet, an increasing number of people start to create and use blogs, which leads to the explosion of sentiment information. Automated techniques for analyzing author's attitudes towards specific events will attribute to business intelligence and public opinion survey because it is hard to gather and process such massive information only with manual methods. The classification methods in the existing blogging sites are almost based on topic and hard to get the information of blogs. Therefore, it's a well-worth concerning issue that how to get the sentiment in blogs.

Text sentiment analysis has had world-wide attention since 1990s. The available sentiment analysis techniques fall into two categories, semantic approach and machine learning.

Mckeown and Hatzivassiloglou have begun to research on semantic orientation since 1997. J.Kamps, R.J.M Okken, M.Marx and M.D Rijke used WordNet to measure the semantic orientation. They chose several pairs of bipolar adjectives to scale the responses of subjects to words, short phrases, or texts. Some clues in dispersive verbs and adjectives have been found, and the praise and criticism in sentences have been measured by N-gram analysis. With the functions of calculating semantic similarity and semantic relevance provided by How Net, Ly Zhu et al. proposed a method to measure semantic orientation, which is to calculate the relevance of candidate words and the seed words. B.Pang et al. employed three machine learning methods (Naive Bayes, maximum entropy classification, and support vector machines) to research on sentiment classification. Lh Xu et al. adopted the derogatory or commendatory terms as features of classification to create Support Vector Machine classifier and identify the text orientation.

Chinese is different from English that there is no clear boundary between its words. Thus, we need Chinese word segmentation techniques to pre-process them. Chinese word segmentation techniques are divided into three categories, each of which is based on string matching, statistics and comprehension respectively. The word segmentation technique based on string matching is simple and efficient. However, it is hard to use this technique to solve the problems in ambiguous recognition and new word recognition. The technique based on statistics doesn't need a word segmentation dictionary but has low accuracy. And the technique based on comprehension is to make machine own human's comprehensive ability but it is still in early stages.

Comment Target Extraction.

The system in this paper uses the sentiment analysis technique based on comment target extraction. The framework of algorithm is shown below. Firstly, we use Pan Gu segment to extract nouns and noun phrases as the candidate comment targets. Secondly, we use term-frequency filtration to filter those targets, dividing the sentences into several categories and analyze sentiment orientation for each category. At last, we compare the quantity of positive and negative sentiment sentences and then we can get the sentiment orientation of the blogs. Fig. 1 is the system flow diagram.

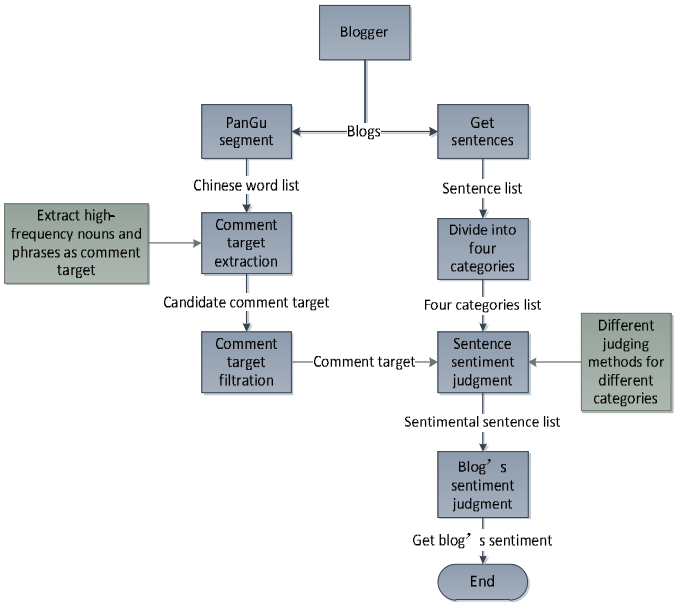

Fig. 1. System Flow Diagram

The sentiment analysis technique in the paper is based on comment target extraction. For given material, we use Pan Gu segment and extract nouns and noun phrases as the candidate comment targets. Then we filter the candidate comment targets by term-frequency filtration and NN filtering algorithm, and finally get the table of targets.

The targets have some noise so we adopt filtration technique. The technological process has two parts. First, use term-frequency filtration to filter the targets. Second, use NN filtering algorithm to filter redundant nouns.

Term-frequency Filtration.

We consider that comment targets are tend to appear in comments and some irrelevant nouns are seldom appeared in wares, such as 'limited company' and 'images'. We also find that term-frequency may filter some comment targets.

NN Filtering Algorithm.

This algorithm mainly deals with the filtration of redundant nouns. In order to explain what redundancy is, we define s-support. For a noun t, assume that there are sentences containing t. In those s sentences, there are k sentences in which appear as individual comment targets (t is not shown in those k sentences). Then s-support=k/s. When the value of s-supports is less than the threshold of some comment target, we regard it as a redundant noun and filter it. The threshold value can be set to 0.5.

Orientation Judgment on Comment Target.

While making orientation judgment on comment targets, we analyze the structure of subjective sentences and divide them into four categories. Then we set different judgment rules for each category. Finally, we judge the blog's sentiment orientation according to the absolute value of the difference of positive and negative sentiment sentences.

Categories of Sentiment Sentences.

Category 1

The sentence has obvious orientation. The context-free words with some kind of orientation (positive or negative) are much more than the other kind.

Category 2

The sentence has no obvious orientation but there are an equal number of positive and negative sentimental words.

Category 3

The sentence has no obvious orientation but its context has obvious orientation.

Category 4

The sentence has no obvious orientation and there is no sentimental word in it. Neither is its context.

Text orientation judgment.

For category 1, the polarity of the comment targets in sentence is same as the sentence's polarity.

For category 2, find out comment targets in the sentence. For each comment target, we select the nearest sentimental word as the directly-modified word within the limit of 8 words (experiment shows this size is suitable) of the comment target.

For category 3, use context information to judge orientation. It's a priority that the current sentence has the same polarity with its previous sentence. If there's no obvious orientation in the previous sentence, we consider the current to have the same polarity with its next sentence. After the orientation judgment of the sentence, the polarity of all comment targets in this sentence is the same as this sentence's polarity.

For category 4, find out comment targets in the sentence. Then find out the nearest context-dependent sentimental word. If the binary pair <sentimental word, comment target> has polarity, we extract the result. Otherwise we filter it. The table of context-dependent sentimental word is small and the word with comment targets can result in orientation. The table is established manually.

Finally, we judge the blog's sentiment orientation according to the absolute value of the difference of positive and negative sentiment sentences.

Testing Results.

We have tested our content-based blogger's sentiment analysis system by using thirteen teacher's blogs. We've also compared the result of machine analysis with the result of manual analysis. The experimental result is shown in table 1.

Table 1. Sentiment analysis testing result

Blog File	Manual analysis results (sentiment classification)	Software analysis results (sentiment classification)	Words count	Events count	Consistency between manual analysis and software analysis
2010.1.2.txt	positive	positive	356	6	Yes
2010.11.11.txt	negative	negative	57	1	Yes
2010.12.14.txt	positive	positive	300	6	Yes
2010.4.1.txt	positive	positive	118	2	Yes
2010.5.24.txt	positive	positive	289	3	Yes
2010.6.18.txt	negative	positive	346	6	No
2010.6.9.txt	negative	positive	1004	29	No
2010.7.15.txt	negative	positive	876	13	No
2010.9.11.txt	positive	positive	263	7	Yes
2011-6-27.txt	negative	positive	237	2	No
2011.3.13.txt	positive	positive	561	4	Yes
2011.3.8.txt	negative	negative	184	2	Yes

The accuracy of the system is 66.6% as it is shown in the table above, which illustrates our system's effectiveness. But we still find a problem that the system has a low accuracy when analyzing blogs with negative sentiment. It has analyzed only one of four blogs with negative sentiment and the accuracy is only 25%. One of the reasons for the problem is that the negative sentimental words in vocabulary are not comprehensive enough.

Conclusions

This paper describes a system for analysis of the blogger's sentiment based on Pan Gu segment component and its related technologies– comment target extraction and comment target orientation judgment. The system has a good test result. Moreover, sentiment analysis technology is still in its infancy and has wide research space.

References

[1] I. Bayoudh, N. Bechet, M Roche. Blog Classification: Adding Linguistic Knowledge to Improve the K-NN Algorithm. IFIP International Federation for Information Processing, (2008)68–77

[2] A. K. Singh, R.C. Joshi. Semantic Tagging and Classification of Blogs. Computer and Communication Technology (ICCCT), 2010 International Conference on, (2010)455 – 459

[3] Geetika T. Lakshmanan,Martin A. IFIP Advances in Information and Communication Technology (2008) 68-77

[4] Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences[C]. In Proceedings of IJCNLP-2005.(2005)61–66.

[5] Jun Zhao, Kang Liu, and Gen Wang. Adding redundant features for crfs-based sentence sentiment classification[C]. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. (2008)117–126.

[6] PANG B，LEE L．Seeing stars：Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics，(2005)1 15-124 EGUCHI K， LAVRENKO V． Sentiment

[7] Erdmann M, Ikeda K, Ishizaki H, et al. Feature Based Sentiment Analysis of Tweets in Multiple Languages[M]//Web Information Systems Engineering–WISE 2014. Springer International Publishing,(2014)109-124.

[8] Rosenthal S, Nakov P, Kiritchenko S, et al. Semeval-2015 task 10: Sentiment analysis in twitter[C]//Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval. (2015)

[9] Paltoglou G. Sentiment analysis in social media[M]//Online Collective Action. Springer Vienna, (2014)3-17.

[10] Cambria E, Song Y, Wang H, et al. Intelligent Systems, IEEE. 29(2014,)44-51.

[11] Poria S, Cambria E, Winterstein G, et al. Knowledge-Based Systems. 69(2014)45-63.

