http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (3): 510-513, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.510.513

Polarity Analysis of Stock Comments in Chinese Based on Word Clustering

[1]Yang Weijie and [2]Zhu Yapeng
[1]School of Computer and Information Engineering,
Beijing Technology and Business University, Beijing, China
[2]The Research Institute of Data Communication Technology, Beijing, China

Abstract: This paper proposed a method of polarity analysis of stock comments in Chinese based on word clustering, to avoid effect of noisy information without emotional tendency. This method mines semantic association between words to construct the concept dictionary in specific areas dynamically and describes the text feature with the concept constructed. The experimental results show that this method can not only reduces dimensionality of feature space, speed up the process of sentiment classification, but also improve the performance of sentiment classification to a certain extent.

Key words: Stock comments, polarity analysis, word clustering, sentinent, classification

INTRODUCTION

Polarity analysis of internet text plays an important role and can be used to diversity areas in our life. And emotional polarity analysis technology will make the computer much efficient and smarter to deal with the data in a short time; the public opinion will be collected timely and exactly.

Few studies relate to sentiment analysis of stock comments, which is a new research area of polarity analysis. Stock comments refer to the stock analysis reviews given by professional experts. If the polarity of this views can be accurately determined, investors can get more clear stock analysis to help them make better decisions to buy or sell, or hold. Compared with other reviews, stock comments have following characteristics (Mo *et al*., 2011), (1) Specific terms: Stock comments have some specific terms which can be used for polarity analysis, for example, bull market or bear market. But there is no sentimental lexicon and general sentimental lexicon can not be used in this area, (2) Most of feature words are verbs: The traditional calculation of the tendency based on quantitative analysis of emotion words, whose performance is not good when it is applied to the polarity analysis of stock comments, (3) Stock comments reviews usually have many words: Most of words are used to descript the current status and these words are noise data for sentiment analysis actually. In addition, the status description may show the inconsistent view to the tendency expressed by the experts. Finally (4) It is no longer a simpler problem of automatically classifying a piece of text as expressing positive or negative sentiment. Three results can be get from sentiment classification of stock reviews, they are bullish (rising trend), bearish (decline trend) and stable.

RELATED WORKS

At present, the research methods of text polarity analysis can be broadly divided into two categories: semantic-based methods (Hatzivassiloglou and McKeown, 1997) and statistical methods (Xu *et al*., 2007). However, these methods do not consider the semantic integrity and can not obtain good results. And stock comment is a new research field of polarity analysis, there are only several research results and ideas. There are three reasons for this problem. First is that, construction of field dictionary is very difficult work, so the statistical method is not feasible for this question. Second is that mode-based analysis method may not be able to result the central point of the stock analysts accurately and structure of text need be involved to improve the emotional classification (Hu and Mo, 2009). Mo *et al*. (2011) proposed a combined approach of polarity analysis of the stock comments. They combined the text structure based methods and the model based methods and got a good result. Hu (2009) established stock view modes to present the polarity of stock comments according to the specificity of stock terms and characters of feature words of stock comments, to judge the polarity of stock views more accurately.

In this paper, the method based on word clustering, removes the noise information and makes the process of the emotional tendency analysis more reasonable.

Corresponding Author: Yang Weijie, School of Computer and Information Engineering,
Beijing Technology and Business University, Beijing, China

THE CONCEPT OF WORD CLUSTER

In the field of text mining, feature selection and reconstruction is the basic of text information processing. a description method of text features based on word clustering is used here. This method aims at construct concept dictionary by machine learning methods dynamically and uses this kind of concept dictionary to describe the characteristics of the text. The method needs no theme dictionary and uses seed words as text features. For a given text, it extracts key words with information theory firstly and then makes the keywords map to feature space and get the feature vector description of text.

In the field of text information processing, three ways are used to generate the concept (Li and Yamanishi, 2003). First, uniting two words to get a new word as a concept; second, making the upper meaning of a word as a concept; third, clustering a number of words with similar meaning to generate class center as the concept. In this paper, the third method is adopted. The seed words are used as class center, clustering words associated with a particular theme. The seed words are used to represent this class, each seed word representing a concept theme.

The seed words are that closely related to a theme and used to represent the class of the synonyms words. For example: explosion (coal mines, bombs, terrorism, suicide, car, gas, death) is a seed word that is used to represent a word class. Each word of this word class is an element of seed words used to represent the word class. In this example, "Mine", "bomb" and "terror" and "suicide" are elements of the word class represented by seed word “explosion”, or referred to as elements the of seed word "explosion".

CHI algorithm is used to calculate the correlation between the words w_i and w_j. If w_i and w_j are two words and complied with distribution of χ^2 with one-order freedom degrees. The greater the χ^2 value of w_i and w_j, the higher the correlation of w_i and w_j Character n is the total number of training corpus. Character a is the number of documents in which w_i and w_j are included together, b is the number of documents in which w_i appears only. Character c is the number of documents in which w_j appears only, character d is the number of documents in which w_i and w_j are not included. The correlation of w_i and w_j can be got by the equation as follows:

$$\chi^2(w_i, w_j) = \frac{n \times (a \times d - c \times b)^2}{(a+c) \times (b+d) \times (a+b) \times (c+d)}$$

Following method is used to obtain the feature words of seed words list from training corpus.

- **Text preprocessing:** The training texts are segment into words and removed stop words and candidate words set $U = \{w_1, w_2 \ldots w_n\}$ is get
- **Seed word selection:** Words with word frequency greater than the threshold f are selected from U to construct the seed words set $V = \{sw_1, sw_2\ sw_m\}$
- **Word clustering:** Each word w_j (j =1,2,..., m) of candidate words set U, is calculated its correlation value $\chi^2(sw_i, w_j)$ with each seed word sw_i (j = 1,2,...,m) of V. If χ^2 is larger than a given threshold r and satisfies the condition w_j will be added into the word class represented by sw_i, Otherwise, the correlation value of next word of U and seed word sw_ineed be calculate, until the same calculation of all words in U are completed. The final class word can be get, denoted by sw_i (w_{i1},w_{i2},...w_{ik}) among which sw_i is the ith seed word, w_{i1}, w_{i2},...w_{ik} are seed word elements of sw_i. Feature of texts can be represented by vectors of seed word as (sw_1, sw_2,..., sw_m)
- **Clculation of weight factor of word class:** Let sw_i (w_{i1}, w_{i2},...w_{ik}) is the word class represented by seed word sw_i, average amount of information contained by this word class is as the weight factor of this word class, calculated by:

$$G(sw_i) = -\frac{1}{k}\sum_{j=1}^{k} p(w_{ij}) \log p(w_{ij})$$

where, G (sw_i) is the weight factor of sw_i, p (w_{ij}) is the probability distribution of jth element of sw_i. Weighting factor reflects the ability of feature items to distinguish the attributes of text categories in the feature space.

TEXT FEATURE REPRESENTATION

For a given text D, feature vector is generated by steps of keywords extraction, feature generation and weight calculation.

Keywords extraction: For the given text, d firstly, keywords are extracted based on Shannon information theory from d (Pang *et al*., 2001). After preprocessing, d can be represented by a word list d = (w_1, w_2,...w_n). Assumed that the text d is generated according to a discrete probability distribution, p(w) where, the value range of random variable w is word list d. The amount of word w_i is H(w_i), calculated as follows:

$$H(w_i) = -N(w_i) \times \log p(w_i)$$

where, N (w_i) is the frequency of word w_i in d, (p (w_i) is the probability distribution of, w_i, calculated by equation:

$$p(w_i) = F(w_i) \times F$$

where, F is the total word frequency of training texts corpus, F (w_i) is the total word frequency of w_i in training texts corpus. Words whose H (w_i) value is greater than a threshold value are selected as keywords.

Feature generation: For a given text d,h keywords extracted are $w_1, w_2, ...w_h$, and then, they are mapped to feature space using the following methods:

- If keyword w_i (i = 1,2,...h) is a seed word, it will directly be mapped into a feature word of text d
- If keyword w_i is not a seed word, but w_i is an element of sw_i, it will be mapped into a feature word represented by sw_i
- If keyword w_i is not a seed word and w_i is not an element of any seed word, it will be removed from feature space

Weight calculation: According to the ways of mapping from keywords to feature space, weight of each feature is calculated.

If keyword w_i of d is, feature word, its weight is $Q(sw_i) = G(sw_i)H(w_i)$, $Q(sw_i)$ is the weight of w_i, G (sw_i), is the weight factor of word list and H (w_i) is the amount of information that wi contains in text d.

If keyword w_i is not a seed word, but $w_i$is a element of sw_i, sw_i, is a feature of text d and its weight is Q (sw_i) = χ^2 (w^i,w^j) G (sw_i) H (w_i) Where, Q (sw_i) is the weight of w_i , χ^2 (w_i, w_j) is correlation value of w_i and, w_j, G (sw_i) is the weight factor of word list and H (w_j) is the amount of information that w_i contains in text d.

However, in the process of feature generation, two or more keywords may generate the same feature word, for example, w_i and w_j are keywords of d, w_i is seed word, w_j is not seed word but w_j is a element of w_i, so the feature words generated from w_i and w_j is w_i. So the feature words of these two keywords should be combined into one feature, whose weight is Q (sw^i_i) = (1+χ^2 (w_i, w_j)) G (sw_i) H (w^i_i).

EXPERIEMENTS AND RESULTS

Classifier: The KNN algorithm is used as the classifier in this paper which is a method for classifying objects based on closest training examples in the feature space. The distance function is defined as follows:

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{(\sum_{k=1}^{M} W_{ik}^2)(\sum_{k=1}^{M} W_{jk}^2)}}$$

Table 1: Result of dimension reduction

	Document set ID		
Factor	1	2	3
No. of original feature words	3545	3326	3587
No. of result feature words	2117	2309	2166

Table 2: Result of polarity classification

	Method			
Characteristics	DF	MI	IG	Proposed method
Precision (%)				
BU	54.5	66.6	69.3	82.7
BE	57.1	65.0	67.5	83.0
ST	43.2	50.4	53.2	72.6
Recall (%)				
BU	79.2.	77.1	79.4	79.8
BE	78.5	75.2	80.3	75.9
ST	50.6	53.6	58.4	73.0
F-measure (%)				
BU	64.6	71.5	74.0	81.2
BE	66.1	69.7	73.3	79.3
ST	46.6	52.0	55.7	72.8

BU: Bullish, BE: Bearish, ST: stable

where, d_i is unlabeled text sample, d_j is training sample,W_{ik} is the feature of the unlabeled text sample and W_{jk} is the feature of the training sample.

Data set and evaluation metrics: The 3000 stock comments are accumulated as the test data corpus, in which, the number of stock reviews sets express three trends are 1000 for each. These reviews are crawled from prominent sites. Three metrics used to evaluate the performance of classification are recall, precision and F-measure.

Experimental results: The result of reducing characteristic dimension In order to verify effect of the proposed method to reduce characteristic dimension, three document sets are get by the way that select 300 documents from test data corpus, 100 documents for each kind of stock comments. And the result set of feature words are obtained by feature structure of the original set of candidate words. The experimental results are shown in Table 1.

The result of polarity analysis: Each type of stock comments are divided into five parts, each part contains 200 documents and then cross-experiment is performed. For each time, one document set is select as the training set, the other set are test sets. The average of five results will get as the final result.For comparison, methods introduced in article (Yang and Qian, 2011) are selected as the baseline method. The results are shown in Table 2.

CONCLUSION

Text characterization is an important part in text classification and it can impact on the effect of text

classification. In this paper, a text feature extraction method based on word clustering is used in polarity classification of stock comments. This method uses the semantic association between words to express the content, eliminate the synonymous, reduce the dimension of feature vector and improve the efficiency of text classification. This method combines with KNN classification algorithm to make the experimental results more satisfying and provides new ideas for how to make better use of semantic information in the text sentimental classification.

ACKNOWLEDGMENTS

This work is supported by the Research Foundation for Youth Scholars of Beijing Technology and business University QNJJ2010-25, Beijing Municipal Organization Department talents project 2011D005003000016, Beijing Philosophy and Social Science Planning Project 11JGC100.

REFERENCES

Hatzivassiloglou, V. and K.R. McKeown, 1997. Predicting the semantic orientation of adjectives. Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics, July 7-12, 1997, Madrid, pp: 174-181.

Hu, H., 2009. Research on Web-based Opinion Analysis for Stock Reviews. Beijing Technology and Business University, Beijing.

Hu, H.L. and Q. Mo, 2009. Research on opinion classification of stock recommendations based on discourse structure. J. Chin. Comput. Syst., 5: 896-899.

Li, H. and K. Yamanishi, 2003. Topic analysis using a finite mixture model. Inform. Process. Manage., 39: 521-541.

Mo, Q., Y.J. Zhang and H.L. Hu, 2011. Combined approach of polarity analysis on stock analyst. Comput. Eng. Appli., 47: 222-225.

Pang, J.F., D.B. Bu and S. Bai, 2001. Research and implementation of text categorization system based on VSM. Appli. Res. Comput., 18: 23-26.

Xu, L.H., H.F. Lin and Z.H. Yang, 2007. Text orientation identification based on semantic comprehension. J. Chin. Inform. Process., 21: 96-100.

Yang, W.J. and H.B. Qian, 2011. Comparative experiments on effects of thesaurus construction for sentiment classification of stock comments in Chinese. Proceedings of the International Conference on Communications and Intelligence Information Security, (ICCIIS'11), Hang Zhou, China, pp: 867-870.