A Hybrid Method of Feature Selection for Chinese Text Sentiment Classification

Suge Wang
School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
School of Mathematics Science, Shanxi University, Taiyuan 030006, China
wsg@sxu.edu.cn

Yingjie Wei
School of Mathematics Science, Shanxi University, Taiyuan 030006, China
majie@sxu.edu.cn

Wu Zhang
School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
zhang@staff.shu.edu.cn

Deyu Li
School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China
Lidy@sxu.edu.cn

Wei Li
School of Mathematics Science, Shanxi University, Taiyuan 030006, China
liweikeai@gmail.com

Abstract

Text sentiment classification can be extensively applied to information retrieval, text filtering, online tracking evaluation, the diagnoses of public opinions and chat systems. In this paper, a kinds of hybrid methods, based on category distinguishing ability of words and information gain, is adopted to feature selection. For examining the impact of varying the feature dimension to classification results, using corpus of car reviews, feature dimensions, 1000, 2000 and 3000 are adopted in our experiments. The experiments classification results indicate that the hybrid methods are best with feature dimension equal to 3000, and the result by using hybrid methods is superior to that by directly using information gain. In our experiments F value can achieve over 80%. Finally, some mistake examples are employed to indicate the limitations of methods in this paper.

1. Introduction

Text classification has become one of the key technologies in processing and organizing large amount of document data. It can effectively resolve information disorder and accurately locate required information. At present, text categorization is usually based on topics, such as economy, education, sports, polity, military affairs and so on[1-3]. However, large amount of texts not only include topic information but also include opinion, attitude, sentiment, emotion, likes and dislikes, etc., such as discussions and reviews to some events, evaluation of some products, attitudes to international relationships, opinions to a sport team and its players, such information hidden in texts are called sentiment information. Text classification by mining and making use of sentiment information is called sentiment classification. Sentiment is a very extensive concept which deals with human views, opinions and attitudes etc. By sentiment category of a text we mean the sentiment orientation (positive or negative) and intensity of the text. Sentiment classification technology can be extensively applied to information retrieval, text filtering, online tracking evaluation, the diagnoses of public opinions and chat systems. In real life, especially in web the style randomicity and content openness of texts challenge the traditional text classification technologies. Usually sentiment classification needs to synthesize multi technologies such as machine learning, statistics, natural language processing, linguistics etc.

In recent years, driving by applications, sentiment classification has been becoming an attractive topic in natural language processing. In March 2004, the American association for artificial intelligence held a symposium in this area, entitled "Exploring Affect and Attitude in Text"[4]. B. Pang et al.[5] adopted three machine learning methods (Bayes method, maximum entropy model and support vector machine method) to sentiment classification for movie views in English. The result indicates that support vector machine method is best. Hatzivassiloglou and McKeown[6] used constraints to automatically construct a log-linear regression model. Aidan Finn et al.[7] investigate the use of machine learning for automatic genre classification and different feature for building genre classifiers and their ability to transfer across multiple topic domains. Anthony et al.[8]surveyed four different approaches to customizing a sentiment classification system to a new target domain in the absence of large amounts of labeled data and compare and discuss their advantage, disadvantages and performance. Ying Ying et al.[9] studied the problem of the emotional meaning tagging of Chinese texts by using the multiple relaxation alternate algorithms. Peter. D. Turney et al.[10] introduced a method for inferring the semantic orientation of a word from its statistical association with a set of positive and negative paradigm words. Wang zhiming et

Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)
0-7695-2874-0/07 $25.00 © 2007 IEEE

al.[11] introduced the attributes of emotional evaluation in the Grammatical and knowledge-base of contemporary Chinese. Lexical emotion tagging is studied by means of both qualitative and quantitative approaches.

In this paper, we identify text sentiment orientation, positive and negative, for car product reviews.

(1) The hybrid feature selecting method based on category distinguishing ability of words and information gain is introduced for sentiment classification.

(2) The impact of feature dimension to classification is considered.

(3) The classification results of using hybrid methods are superior to only using information gain and its F value can achieve over 80%.

(4) Some mistake examples are employed to indicate the limitations of methods in this paper.

2. The method of feature selection

In traditional topic classification, candidate features are selected from the residual words after eliminating the stop words from texts. However, for sentiment classification, the words with category distinguishing ability and sentiment orientation should be selected as features. In order to select good feature, we would adopted the hybrid method based on category distinguishing ability of words and information gain. For selecting words with stronger distinguishing ability, five kinds of schemes are attempted. The following abbreviative symbols are used in this paper.

P — positive

N — negative

PIW — the number of training documents, with feature w, belonging to P category

NIW —the number of training documents, with feature w, belonging to N category

PEW —the number of training documents, without feature w, belonging to P category

NEW —the number of training documents, without feature w, belonging to N category

M — the number of training documents belonging to P category

N — the number of training documents

It is clear that N-M is the number of training document belonging to N category, $PIW+PEW=M$ and $PIW+NIW+PEW+NEW=N$.

Scheme 1: If the distribution of a feature w is mean between positive and negative texts it can be considered that no effect to classification is contributed by the feature. Information entropy can be used to depict the distribution of a feature w. By computing the information entropy of the segmented words in training corpus, the words with greater entropy are eliminated and residual words are regarded as features.

The formula of entropy is defined as follows:

$$H(w) = -(p(P/w)\log p(P/w) + p(N/w)\log p(N/w))$$
$$\approx -\left(\frac{PIW}{PIW+NIW}\log\frac{PIW}{PIW+NIW} + \frac{NIW}{PIW+NIW}\log\frac{NIW}{PIW+NIW}\right) \quad (1)$$

Scheme 2: In general, lower frequency words have no contribution to classification. So the words with frequency DF > 1 are selected as features.

Scheme 3: Combining the schemes 1 and 2, i.e., eliminating the words with frequency DF=1 and greater entropy, the residual words are considered as features.

Scheme 4: A feature is called category feature if it appears into the majority of texts of some category. Inversely, it is called a discriminating feature of the category[3]. Hence, one word that is a category feature for some category and is a discriminating feature for another category should be selected as a feature. In this paper, this idea is depicted by defining the following frequency difference.

$$FD(w) = (p(w|P) - p(w|N))^2$$
$$\approx \left(\frac{PIW}{PIW+PEW} - \frac{NIW}{NIW+NEW}\right)^2$$
$$= \left(\frac{PIW}{M} - \frac{NIW}{N-M}\right)^2$$
$$= \left(\frac{PIW\cdot NEW - NIW\cdot PEW}{M\cdot(N-M)}\right)^2 \quad (2)$$

Noticing that both M and N-M in formula (2) are constants independent on the feature w, frequency difference can be simplified into formula (3).

$$FD(w) = (PIW\cdot NEW - NIW\cdot PEW)^2 \quad (3)$$

The more $FD(w)$ is, the stronger the distinguish ability of candidate feature w is.

Scheme 5: Fisher discriminant criterion is a very useful method in machine learning. Its main idea is to make the between-class distance as large as possible and the within-class distance as small as possible when a measure is designed for classification problems [3].

In this paper, the Fisher discriminant is defined as follows:

$$FD(w) = \frac{(E(w/P) - E(w|N))^2}{D(w|P)D(w|N)} \quad (4)$$

where $E(w|P)$ and $E(w|N)$ denote the conditional means of the feature w with regard to the categories P and N respectively, $D(w|P)$ and $D(w|N)$ denote the variances of the feature w with regard to the categories P and N respectively, $(E(w|P)-E(w|N))^2$ and $D(w|P)\cdot D(w|N)$ formulate the scatter degrees between and within the categories P and N respectively.

The more the ratio $\frac{(E(w|P)-E(w|N))^2}{D(w|P)\cdot D(w|N)}$ is, the stronger the distinguish ability of w is.

Fisher discriminant (4) can be simplified as formula (5)

$$FD(w) \approx \frac{(PIW \cdot NEW - NIW \cdot PEW)^2}{PIW \cdot NIW \cdot PEW \cdot NEW} \quad (5)$$

Introduction of information gain refers to literature [2].

3. The process of text sentiment classification

Tentative experiments in advance indicate that information gain is superior to mutual information and X2 statistics for feature selection, and Boolean weight is superior to frequency weight. The experiment results coincide with the existing reports about sentiment classification in English [5]. The whole experiment process is divided into training and testing parts.

(1) Segmented preprocess of training texts.

(2) Select classification features by using category distinguishing ability and information gain.

(3) Express each text in the form of vector by using Boolean weights of features.

(4) Train the support vector classification machine by using training data.

(5) Test the performance of the classifier by using testing data.

Introduction of support vector machine refers to literature [3].

4. Corpus and experiment result

4.1. Corpus selection

The corpus of car reviews is collected from the web as the experiment data in this paper. The documents are then manually labeled with positive or negative symbol. The car reviews, with 337 positive and 143 negative cases respectively and total 420 thousands words, were published from June to August in 2006. 11 kinds of car trademarks were reviewed in this corpus.

4.2. Experiment result

In this paper, the criteria of system performance evaluation contain the recall, precision and F value. The evaluations were for positive, negative and total documents respectively. We performed the experiment in five-fold cross validation.

Experiment: Two ways of feature selection are considered in this experiment. One is only based on information gain. And another is based on both distinguishing ability of words and information gain, we called it hybrid method. According to Section 2, the category distinguishing ability of a word can be measured with 5 schemes respectively. So there 5 approaches (Scheme i + information gain, i = 1, 2, …, 5) should be considered. Feature dimensions, 1000, 2000 and 3000 are selected in this experiment for checking the impact of feature dimensions. The experimental results are shown in Figure1 and Table 1.

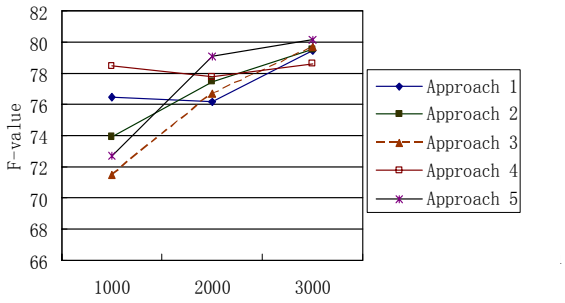

Figure 1. The effect varying the number of dimensions for approach 1-5

Table 1. The results of experiment

Feature Dimension	Document Type	Evaluating Criterion（F-value）					
		Approach					Information Gain
		1	2	3	4	5	
1000	Positive	83.18	79.76	78.97	84.20	78.98	84.40
	Negative	62.10	65.13	56.28	66.98	62.00	67.11
	Total	76.46	73.96	71.46	78.46	72.70	78.60
2000	Positive	82.45	83.35	80.17	83.56	84.60	78.32
	Negative	63.72	66.42	64.51	66.75	67.95	57.44
	Total	76.19	77.48	76.70	77.77	79.10	68.55
3000	Positive	84.59	84.73	84.78	84.04	85.27	81.62
	Negative	70.14	70.78	70.28	68.96	70.84	63.53
	Total	79.46	79.52	79.72	78.60	80.18	75.20

In Figure 1, the F values of sentiment classification of method 2, method 3, and method 5 increases as the dimensions of feature increases. However, the curves of F values of method 1 and method 4 look quite noisy. With 3000 dimension, the largest F value of method 1 to method 5 is achieved and the F value of method 5 is largest compare than others.

For five kinds of hybrid approaches, under 3000 dimension, Approach 4 tends to do the worst and Approach 5 tend to do the best, although the differences aren't very large.

In Table 1, the F values of positive documents are larger than that of the corresponding negative documents. The F values of the hybrid methods are larger than that of method only based on information gain under 2000 and 3000 dimensions. As a whole, the classification result becomes better when the words with strong distinguishing ability are used as classification features. It means that the hybrid feature selection method is effective for text sentiment classification.

5. Discussions

The experiment results indicate that the hybrid feature selection method for sentiment classification is quite good in comparison to the method only based on information gain. However, we were not able to achieved F value on the sentiment classification problem comparable to those reported for standard topic-based categorization, despite the several different types of methods of feature selection we tried. The following observable phenomena would be the reason.

(1) By the presented method, although some words depicting sentiment orientation, such as "serious(严重), advanced(先进), joy(喜悦), jerry(偷工减料), strong(强烈), mild(轻微), durable(耐用), reluctance(勉强), rapid(快速), gorgeous(华丽), fragile(脆弱), slap-up(高档), stimulate(刺激), satisfaction(满意), secure(安全)" and so on, can be selected as classification features, other words, without close relation to sentiment orientation, such as "distance(距离), GPS, highway(公路), gearlever(变速杆), Santana(桑塔那)" and so on was also selected as classification features, and the classification accuracy may be decreased.

(2) In the real corpus, like Corpus adopted in this paper, the size of negative documents is smaller than that of positive documents. It leads to a small F value of negative documents, and so that a small F value of whole documents. Another experiment shows that the classification result can be improved for a corpus with balanced positive and negative documents.

(3) The software of segmentation and part of speech tagging is used and the result is not checked by manual. Made some error of segmentation and part of speech tagging were used, such as, car Brand "Peugeot(标致)" is tagged adjective.

6. Conclusions

Wherever Times is specified, Times Roman, or New Times Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times that you have access to. Please avoid using bit-mapped fonts if possible. True-Type 1 fonts are preferred. Text sentiment classification can be extensively applied to information retrieval, text filtering, online tracking opinions in online discussion, analysis of survey responses, and chat systems. In this paper, a kind of hybrid methods, based on category distinguishing ability of words and information gain, is adopted to feature selection. For corpus of car reviews, the classification result by using hybrid methods is superior to that only using information gain. In our experiments, F value can achieve over 80%. However, the F value of the sentiment classification is lower than those reported F value for standard topic-based classification. It shows more difficulty to the feature extraction for sentiment classification problems. Our further research works will focus on establishing a sentiment knowledge base based on vocabulary, syntactic and semantic and ontology.

Acknowledgments

This work was supported by the National Natural Science Foundation No.60573074, Natural Science Foundation of Shanxi Province No.20041040, Shanxi Foundation of Tackling Key Problem in Science and Technology No.051129, Science and Technology Development Foundation of Colleges in Shanxi Province No.200611002 and Natural Science Foundation of Shanxi Province No.2007011042. We also would like to thank Prof. Maosong Sun for his many valuable comments.

References

[1] D.J. Xue, M.S. Sun. "A study on feature weighting in Chinese text categorization", Computational Linguistics and Intelligent Text Processing (CICLing-03), LNCS2588, Springer-Verlag, pp. 592-601 , 2003.

[2] R.L. Li, "The Key Techniques Research on Text Categorization", Ph.D. dissertation Shanghai: Fudan University, 2005(in Chinese).

[3] F.X. Song, D.W. Zhang, J.Y. Yang and X.M. Gao, "Adaptive classification algorithm based on maximum scatter difference discriminant criterion", Acta automation sinica. 32(4), pp. 541-549 , July 2006 (in Chinese).

[4] B.Philip and T. Hastie, S. Vaithyanathan, "The Sentimental Factor: Improving Review Classification via Human-Provided Information", The 42nd Annual meeting

of Association for computational linguistics, pp. 263-270, 2004.
[5] B. Pang.L, Lee and S.Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 79-86, 2002.
[6] V. Hatzivassilolou Kathleen, R. Mckeown, "Predicting the semantic orientation of adjectives", Proceeding of the 35th Annual meeting of the association for computational linguistics and the 8th conference of the European Chapter of the ACL, Association for Computational Linguistics, New Brunswick, pp. 174-181, 1997.
[7] A. Finn, N. Kushmerick and B. Smyth, "Learning to classify documents according to genre", Journal of the American society for information science and technology(JASIST), Special Issue on computational analysis of style, 7(5), pp. 1415-1562, March 2006.
[8] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: a Case Study", In Proceedings of RANLP, 2005.
[9] Y. Ying, F. Zhou and C.L. Zhou, " A Research on Emotion Tagging of Chinese Understanding by Designing an Experiment System", Journal of Chinese Information Processing. 16(2), pp. 27-33 , 2002. (in Chinese)
[10] Peter D. Turney Michael L. Littman, Measuring "Praise and Criticism: Inference of Semantic Orientation from Association", ACM Transaction on information systems, 21(4), pp. 315-346, 2003.
[11] Z.M. Wang, X.F. Zhu and S.W. Yu. "Research on Lexical Emotional Evaluation Based on the Grammatical Knowledge-Base of Contemporary Chinese", Computational Linguistics and Chinese Language Processing, 10(4), pp. 581-592, 2005. (In Chinese)