Research on recognition algorithm of network public opinion in view of evaluation

Chenxiang Zhang

Suzhou Industrial Park Institute of Services Outsourcing, Suzhou Jiangsu 215123, China

kylinbaby@163.com

Keywords: Internet public opinion; Online review; Opinion analysis; Feature presentation; Ensemble learning; Combination optimization; Review spam detection

Abstract. In This paper, firstly, discusses the theoretical background, the existed research achievements and commercial products of Internet public opinion monitoring and analysis, according to preceding context we find the needs of public opinion analysis and the deficiency in the existed systems, and then raise the prototype of Internet public opinion mining with the function of opinion tracking. Secondly, decomposes the prototype to figure out the technology points, then corroding to them, gives a detailed introduction of their theories .Thirdly, describes the overall system design and detail works in each module of the Internet public opinion monitoring and analysis system.

Introduction

Now the social and the Internet environment is more and more complicated. In this situation, the Cyberspace public opinion has already caused important influence on people who are surfing on the Internet. The Cyberspace public opinion is different from the traditional one. In regard to the Cyberspace public opinion, there occurrence range is wide, spreading speed is high, and their eruption spot are difficult to be detected and controlled [1]. All above situation proved that: it is more important to effectively detect and control the Cyberspace public opinion.

Recent years, opining mining research is very active [2]. It is belong to the scope of natural language analysis. Its main purpose is to classify a document into negative or positive based on the content text. In this issue, we analysis the characteristics of the Cyberspace pubic opinion base on the theories of the opining mining.

With the development of Internet technology and applications of network, the Internet has become an important source， even the main source from which public access to information. It also becomes an important place for people to exchange information and to express their views [3]. Understanding Internet public opinion through the net and concerning about trends of Internet public opinion have important practical significance for the maintenance of social harmony and stability and also for the promotion of social democracy and legal construction.

Information on the Internet is so enormous that the identification and assessment in manual way are powerless. How to use computer network technology, artificial intelligence techniques and data mining techniques to mine and analyze the Internet public opinion becomes a new research focus [4]. There are some urgent and important issues in this domain, such as how to identify and categorize the hot topics, from public opinion information on web; how to determine whether the attitude of the people on an social event is positive or negative; how to analyze trend of the fluctuation of the hot social events, etc. These issues have important scientific and practical significance on recognizing and guide of Internet public opinion.

Organization of the Text

This dissertation does some research in the mining and analysis of Internet public opinion information, such as utilizing Web document classification technique to do some classification of emergencies on Web, adopting machine learning methods to analyze the sentimental orientation of

© 2015. The authors - Published by Atlantis Press

Internet public opinion, and giving statistic analysis for fluctuation in investigating the trend of Internet public opinion.

1. In this dissertation, Fisher discriminate analysis is introduced in text classification of Internet public opinion, and then so is the classification of emergencies. Internet public opinion, caused by the unexpected events, is in the content form of document, classification of Internet public opinion is then converted to one text classification problem [5]. Fisher criterion is an effective way in solving the dimensionality reduction problem, but few studies are available in text classification. As a feature extraction method, Fisher discriminate criterion is applied for text classification problems, and then is used in classification of emergency management. As for Internet public opinion research, according to accordance public safety, emergencies are always classified into four types of emergencies, sudden natural disasters, accidents, disasters, public health events, social security events. The experiments proved that the Fisher criteria slightly inferior to the method of information gain, but compared with other feature extraction methods are better. (Fig. 1)

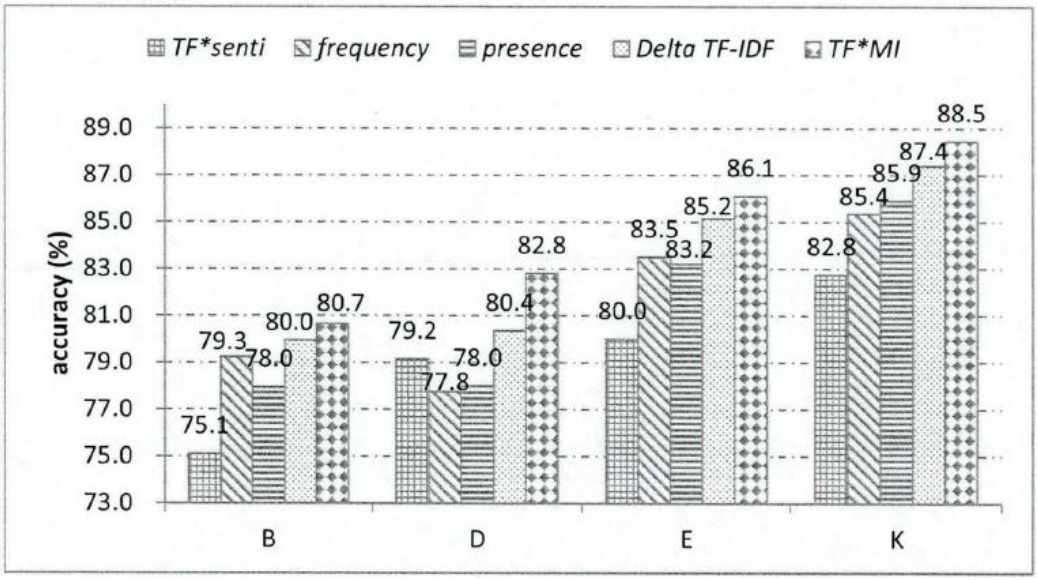

Fig. 1 Unigram type

2. Through analyzing of latent semantic analysis theory, including the singular value decomposition, calculation of the similarity relations between documents, this dissertation proposed a new algorithm, LR-LSA, for web texts classification. In order to eliminate the analysis of the limitations of the method of latent semantic analysis, the proposed algorithm LR-LSA, SVM classifier for a category relevance of each document, and then uses the correlation to different categories to generate the local area. Two classification experiments on Chinese corpus verified that the performance of the LR-LSA is more effective than LSA. (Eq. 1) [6]

$$f(X) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon \tag{1}$$

3. The method based on machine learning for the sentiment analysis considers less sentimental feature extraction, and then in this dissertation we present a method, PMML, combining the feature extraction and machine learning for analysis of sentiment. We introduced relevant research, including the classification at the different granularity, and the methods based on machine learning. Web comments texts, after conducting basic sub-word first, are divided into set of keywords. For those key words, we designed some patterns which are often used in emotional expression. After the success matching to those patterns, emotional features are gotten and in the form of sequence. For each feature pattern, we calculated separately the emotional tendency, and then adopting machine learning method finally to obtain the emotional tendencies of the web comment. The experiments illustrate the effectiveness of the PMML when compared to machine learning method in the classification performances. (Eq. 2)

$$f(X) = \beta_0 + \sum_{i=1}^{p}\sum_{j=1}^{q}\beta_j x_i^j + \varepsilon \tag{2}$$

Method

Compared with traditional text analysis tasks, opinion analysis is more difficult due to the flexibility and complexity of opinion expression. The informal review contents with huge data volume bring more new challenges. On the other hand, opinion analysis contains more research contents, it includes the quality control of opinion texts, opinion information extraction, opinion identification, opinion summarization and retrieval, which is penetrating from data collection and integration to providing analytical results for the users and Web services in next stage. The quality control of opinion texts provides reliable data for the subsequent applications and researches, the opinion identification provides the important information for opinion summarization and retrieval in the opinion analysis process. Therefore, these two research contents are focused on in this paper, and the major contributions are as follows.

1. Proposing a new feature function via integrating the term's sentiment information with its contribution to a document. The traditional feature functions work with poor effectiveness for sentiment classification tasks, because they have not considered the terms' sentiment. The effectiveness of sentiment classification would be promoted by taking advantage of the terms' sentiment. At first, a term's sentiment orientation is captured by evaluating the term's mutual information with the sentiment labels. Secondly, the feature's value is determined by integrating the term's sentiment score and its contribution to a document. The experimental results show that the proposed method is more effective than the traditional ones in sentiment classification. (Eq. 3)

$$P(Y = K|X) = \frac{1}{1 + \sum_{k=1}^{K-1}\exp(W_k^T X)} \tag{3}$$

2. Proposing a three phase framework for sentiment classifiers are selected to make sentiment classification by which a set of predictions and these predictions are integrated via ensemble learning. A quality evaluation criterion for a set of classifiers is proposed based on classifiers' accuracies and diversity, which can determine a set of optimal classifier used for assemble. A stacking-based ensemble learning algorithm is devised to integrate multiple predictions generated by the selected classifiers. The proposed method outperforms the best traditional single classifier method in different domains. (Eq. 4)

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\zeta_i, \zeta^*) \tag{4}$$

3. Proposing a greedy algorithm for classifier set selection, which solve the problem of combination explosion encountered during the, process of classifier set selection. At first the classifier set selection is transformed into an optimization problem. Secondly, a greedy algorithm is devised to select a set of classifiers based on the candidate's accuracies and diversity. The greedy algorithm is proved to be 2-approximation, which guarantees the quality of selected classifiers. Moreover, the proposed algorithm's time complexity is the number of optional classifiers; which enhances the availability of the three phase ensemble learning framework for sentiment classification significantly. (Eq. 5)

$$\begin{cases} w^T\phi(x_i) + b - y_i \le \varepsilon + \zeta_i^*, \\ y_i - w^T\phi(x_i) - b \le \varepsilon + \zeta_i^*, \\ \zeta_i, \zeta_i^* \ge 0, i = 1, \cdots, n \end{cases} \tag{5}$$

4. Proposing six features for review seam detection based on modeling the review contents and reviewer's behaviors, and devising supervised and unsupervised online review seam detection algorithms respectively. The proposed algorithms can identify the review spam in time, which can not

be done by the existing methods. Furthermore, high identification precision and recall on review span can be achieved by the proposed algorithms. Especially, the unsupervised algorithm can obtain good effectiveness without labeled samples. (Eq. 6 and Fig. 2)

$$f(X) = \underset{Y=C_h,C_l}{\operatorname{argmax}}(P(Y) \prod_{j:x_j=1} P(x_j = 1|Y)) \tag{6}$$

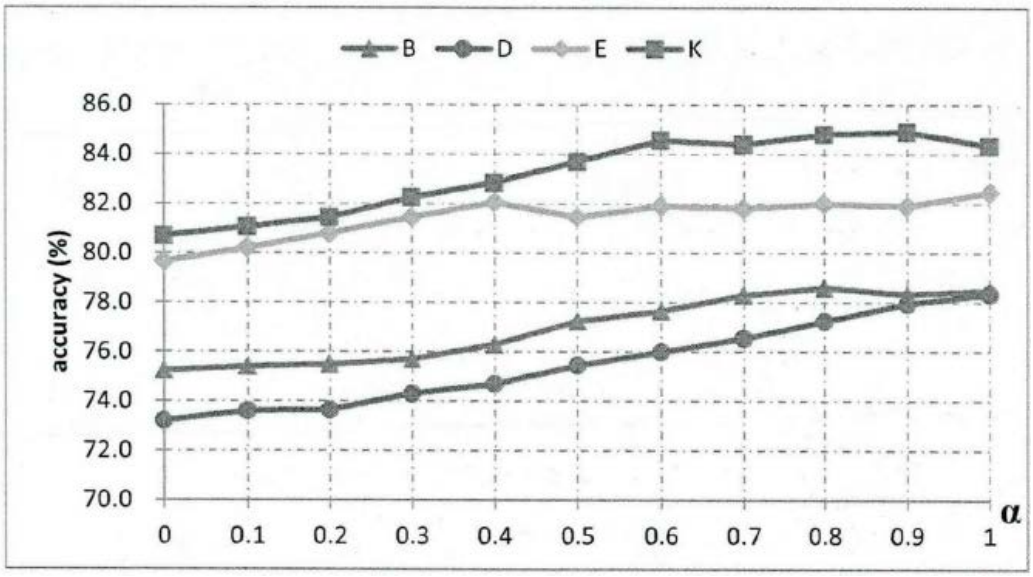

Fig. 2 Influence factors

Summary

In summary, four problems including the feature presentation for opinion analysis, multiple classifiers ensemble learning for opinion analysis, the strategy of classifier set selection, and the online review span detection are studied. These research contents have coherence and sustainability, and form a relatively complete research. Our work is based on the comprehensive survey and analysis on existing theories and techniques. The theory analysis and extensive experiments show that the proposed methods for the four problems above achieve good effectiveness.

References

[1] Bo Pang, Lillian Lee. Opinion mining and sentiment analysis. Foundations and trends in information retrieval. 2 (2008) 1-135.

[2] Minqing Hu, Bing Liu. Opinion extraction and summarization on the web. AAAI. 7 (2006) 1621-1624.

[3] Sihong Xie, Guan Wang, Shuyang Lin, Philip S Yu. Review spam detection via temporal pattern discovery. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, 823-831.

[4] Arjun Mukherjee, Bing Liu, Junhui Wang, Natalie Glance, Nitin Jindal. Detecting group review spam. Proceedings of the 20th international conference companion on World Wide Web. ACM, 2011, 93-94.

[5] Information on http://www.cnnic.cn/ research/

[6] Zakoian, J. M. Threshold Heteroseedastic Models. Journal of Economic Dynamics and Control. 18 (2013) 931-955.