The analysis of online users' emotions based on data mining

Yi Zhou[1, a]

[1]School of Xiangtan University, Hunan411105, China.

[a]zhouyi_0216@163.com

Keywords: social network, sentiment analysis, data mining

Abstract. With the rapid development of social network, analysis for online users' emotions and tendencies has gradually developed. Sentiment analysis and social relations analysis on social network has made significant progress. Based on data mining, this study summarized the existing work about users' emotion analysis, and summarized three sentiment analysis methods of online users. Furthermore, compared the differences of sentiment analysis between Chinese and English.

Introduction

With the rapid development of internet, the way people express emotion and opinion has been widely expanded. The mode of using network changedfrom passive accepting information to initiative expressing their views. More importantly, social network links people through social tie, this can make the information and viewpoint in social interaction to interact and spread, make social network become the medium of users' emotions and viewpoint. In this way, how to find useful information and the rules of users' tendencies in the vast amounts of user generated information is becoming more and more important in further research.

The existing work has explored, analyzed and predicted the emotions and tendencies of online users in many ways.Sentiment analysis can dig out the user's emotional attitude through the analysis of the network text, and can be used in many ways like topic detection, public opinion monitoring and products forecast.And more deeply, the emotions reflected on social network are restricted and influenced by users' character orientation. In return, network information and emotions become the external performance of deeper personality tendency, dig latent layers of personality tendency can help to understand the express of users' emotions and idea.

Sentiment analysis and data mining

The personality tendency analysis for Web text can be divided into four categories, sentiment analysis, topic detection, and public opinion monitoring and user personality tendency analysis. Sentiment analysis also can be calledopinion mining, by using of natural language processing, text analysis and computer linguistics to identify and extract the source of subjective information in information materials. In general, the purpose of sentiment analysis is to dig out the attitude, the context polarity and the emotional state of online users on a particular topic in the Web text.

The classification problem about subjective and objective of web text is the foundation of emotional polarity judgment and strength judgment. The subjective and objective text classification research has already launched both at home and abroad, and applied to the field of information retrieval and information extraction. In the categorization of subjective and objective in English text, the mainstream method is machine learning algorithms. Yu H and Hatzivassiloglou V used similarity method, naive bayesian classification and classification method of multiple naive bayes algorithm to identify the subjective, the accuracy reached 80% - 90% [1]. Pang B and Lee L mainly considered the emotional connection between sentences, used the minimum cuts to improve the classification accuracy [2]. In the categorization of subjective and objective in Chinese text, due to the particularity of Chinese context, mainly used semantic method for judgment.

For emotional polarity analysis of Web text, there are subject matter independent emotional polarity judgment and subject matter related emotion judgment, the differences of those two is the

subject matter related emotion judgment needs to identify the theme, then emotional polarity judgment can be used on each topic. The latter is mainly used for long text, the short text usually unrelated to the topics. The emotional polarity analysis mainly used analysis method based on dictionary and machine learning. The method based on machine learning mainly used the frequent pattern mining method, select different machine learning to judge. The method based on dictionary is mainly based on corpus to build emotional dictionary. Tueney defined the point of mutual information between two specific words (Pointwise MutualInformation, PMI), using PMI to calculate the relevant difference of basic phrases in the text the positive and negative emotional words, to judge the emotional polarity [3].

$$\mathrm{PMI} = \log_2\left(\frac{p(w_1 w_2)}{p(w_1)p(w_2)}\right) \quad (1)$$

Sentiment analysis methods

During the development of sentimental analysis, many effective methods are processed, including vocabulary scale method, emotional dictionary statistics method and statistical learning method based on natural language processing, those are three major methods.

Vocabulary scale method. Vocabulary scale method is the most traditional sentiment analysis method, it comes from the semantic distinction of different cultural evaluation in psychology research. By description of the structure of individual emotional experience, the positive affection and negative affection can be used as two relatively independent dimension [4]. Although vocabulary scale method is not specific for Web text sentiment analysis, as traditional sentiment analysis method, it provides a theoretical basis for Web text sentiment analysis.

Emotional dictionary statistics method. The emotional dictionary statistics method is a sentiment analysis method for Web text based on the traditional vocabulary scale method. The foundation of this method is the emotional dictionary based on corpus, every emotional words in the dictionary is marked by corresponding polarity and strength. There has different ways to build emotional dictionary, for a particular field of sentiment analysis, usually build a typical emotional dictionary for that field. However, emotional dictionary cannot cover all emotional words in the web text because many new network vocabulary contained different emotions emerged every day. What's more, emotional dictionary cannot handle complex sentence structure, it can produce inaccurate analysis on sentence level emotional judgment.

Statistical learning method based on natural language processing. Statistical learning method based on natural language processing usually used frequent pattern mining method, extracted the feature of text, then used the feature to training model, and selected different machine learning methods for judgment. The key point of this methods is extracting meaningful characteristics through the study and digging of a large number of corpus.In specific applications, Pang and Li for the first time used machine learning methods in chapter scale emotion classification task [2].because of the statistical learning method is characteristic-based, the result mainly depends on the features of effectiveness, the effect of combinatorial optimization and feature selection. In the research of statistical learning method based on natural language processing, the main focus is the discovering of effective features and feature selection, feature fusion, and the choice of machine learning methods.

Differences of sentiment analysis between Chinese and English

The analysis of emotion based on Chinese Web text started later than the analysis based on English Web text. The main difference between Chinese web text and English web text is that the former needs text segmentation. Chinese word segmentation refers to a process that cut a sequence of Chinese characters and divide into individual words. The existing of Chinese word segmentation is mainly because the particularity of Chinese basic writing grammar. Firstly, different with the Latin represented by English, Chinese sentence didn't have natural space delimiter, and didn't have

participles writing grammar habits. Secondly, in Chinese sentence, there is no clear boundaries between words, particular case needs particular analysis.

Chinese word segmentation mainly based on three principles: based on dictionary string matching, based on the understanding of syntactic and semantic, based on the frequency statistics.Different segmentation tools use different word segmentation principle or participle combination.Although the existing Chinese word segmentation tools able to input the Chinese text segmentation,but due to the complexity of Chinese sentence semantic, the existing of ambiguities and words the segmentation tools failed to identify, there also can lose meanings and appear error.

The previous section introduced two methods of sentiment analysis, the emotional dictionary statistics method and statistical learning method based on natural language processing, both English and Chinese web text sentiment analysis can use this two method. But in the specific application scenario and effect, there are some differences.

Firstly, for long text, needs to subjective and objective text classification before the study on emotional polarity of subjective texts. The embodiment of the subjective and objective expression of English text is mostly on sentence structure, while the subjective and objective expression of Chinese text more embodies in semantics. Secondly,for short text, it needn’t to subjective and objective text classification before the study on emotional polarity of subjective texts. For Chinese text, a short text could include more content, more statements and the expression of emotion could be more abundant and complicated, the effects of sentence patterns and combination to the emotional polarity. Compared with the sentiment analysis of English text, the accuracy of Chinese text is only about 90%. The collected of corpus, the build of emotional dictionary, the construction of a text feature extraction and the machine learning algorithm are needed to be more in-depth study.

Conclusion

Based on the existing research about online users’ emotions, this study introduced the theoretical basis and research methodology of Web text sentiment analysis, given subsequent research work a sufficient theoretical foundation. By using data mining to analyze and forecast the social network user's emotional tendency, can effectively reveal the relationship between user’s behavior and the user's psychological characteristic in real life. Furthermore, enhance people's understanding of social network.

However, the emotional tendency analysis of online information also has its difficulties. The emotions and its tendencies are the hidden layer characteristics of people, while the information on the social network is mostly the surface layer characteristics; in the perspective of data distribution, compared with the scale data, these online experimental data has more complex structure and lots of noise.Therefore, new innovative approaches are needed to predict online users’ emotion and tendency.

Reference

[1]Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences [C] // Proc. of the 2003 Conf. on Empirical Methods in Natural Language Processing. Sapporo, Japan, 2003:129-136.

[2]Pang, B. & Lee, L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, in 'In Proceedings of the ACL', 2004: 271—278.

[3] TURNEY P. Thumbs up or thumbs down? : Semantic orientation applied to unsupervised classification of reviews [C]//Proceedings of the doth Annum Meeting of the Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, 2002: 417—424.

[4] WatsonD, WieseD, VaidyaJ, etal. The two general activation systems of affect: structural

findings, evolutionary considerations, and psychobiological evidence. Journal of Personality and Social Psychology, 1999, 76(5):820-838.