Topic-Independent Chinese Sentiment Identification from Online News

Zhong-Yong Chen, Department of Information Management, National Taiwan University, Taipei City, Taiwan
Wen-Ting Li, School of Computer Science, University of Birmingham, Birmingham, UK

ABSTRACT

In this paper, the authors investigate the topic-independent Chinese sentiment identification problem from online news. They analyze the word usage and sentence structure of the documents for inferring representative terms and sentences in the documents, and then employ the feature values of each document for identifying the opinion of the topic-independent online news. The support vector machine (SVM) is leveraged for training the classified model in terms of the extracted features and identifying the opinion orientation of the topic-independent documents by the trained model. Experimental results demonstrated that the authors’ features are helpful for identifying the opinions of the topic-independent documents, and can help readers for filtering out the negative documents.

KEYWORDS

Chinese Sentiment Identification, Feature Extraction, Opinion Mining, Support Vector Machine

1. INTRODUCTION

With the explosive growth of Chinese users of the Internet, there are astronomical websites that provide the Chinese news and Chinese articles. The enormous information will be overwhelming for the Chinese readers. However, news always contain different opinion (e.g., positive opinion or negative opinion) (Godbole et al., 2007), that is, news articles not only convey the information people would like to know but also the sentimental components which may influence the emotions of the readers (Russell 2009) and may also affect the productivity of the readers (Wong and Law 2002). Filtering out the news with bad emotions (negative) become an essential task for the readers who only want to know the global trends without affecting performance of their daily works.

However, Chinese news contain a variety of different topics ranging from the business trends to the global warming issues. Investigating the sentiment of topic-independent Chinese online news is much difficult from the sentiment analysis within the same topic in many aspects. First, language-specific characteristic of the Chinese is that there is no delimiter within the Chinese words. When processing Chinese, the information need to be segmented at the beginning that will be affected by the accuracy of the segmenter. Second, Chinese sentence may not comply with the restrict grammar which makes the articles containing more complicated sentence difficult for analyzing the word usage. Finally, the online news covering a variety of topics containing different word usage makes the sentiment analysis of the topic-independent documents (the online news) more difficult.
In this paper, we develop a framework to deal with the sentiment detection of online topic-independent Chinese news. The framework consists of two components: feature selection and sentiment classification. We propose new features to represent the sentimental documents, and employ the support vector machine (SVM) to predict the document sentiments.

The rest of the paper is organized as follows. Section 2 gives the related literatures to our work. In Section 3, we detail each feature we used, and Section 4 provides experimental results to show the effectiveness of our features. Section 5 gives a conclusion and future works.

2. RELATED WORK

With the emerging of Chinese online users, there are astronomical Chinese resources available on the Internet, such as Chinese product reviews, Chinese online news, Chinese forum comments, and Chinese posts from the social media. Filtering out the negative influential information can help readers to improve their productivity (Russell 2009) which the techniques are well-known as opinion mining. Opinion mining, also known as sentiment analysis, has been investigated for decades (Liu 2012; Feldman 2013; Medhat et al., 2014). Ye et al., (2005) investigated the difference between semantic approach and SVM for identifying the opinion of Chinese reviews. Wang et al., (2007) investigated term features by using information gain, entropy and term distribution for identifying opinion from Chinese product reviews. Zagibalov and Carroll (2008) proposed a Chinese sentiment identification method which employed seed words to iteratively extend opinionated word list, and used the list to identify the polarity of the document. Zhang et al., (2009) proposed a two-phase algorithm to predict the sentiment of documents. First, they calculated the sentiment score of each sentence in a given document based on the dependency rules. The word polarity score is based on the HowNet dictionary, and they also took the context (word dependency) into consideration for enhancing the score. Second, they aggregated the scores to represent the sentiment of the document in terms of the importance of each sentence in the document. The importance is calculated by linear combination of five features for a sentence, which the weight of each feature is obtained from the training process. Moreover, Ku et al., (2006) proposed three methods for understanding the opinions: opinion extraction, opinion summarization, and opinion tracking. Opinion extraction method employed the Chinese opinionated dictionary to measure the sentiment score of each word, and then infer the sentiment score of the sentence and the document based on the word-level score. Opinion summarization is based on the importance of word to select the representative sentences to organize the opinionated summary, and the opinion tracking system is a system that represents the opinion changes as time goes by. Due to the scarce of Chinese resource, Wan (2008) proposed a method which translates Chinese into English, and predicts the polarity of words based on the dictionary in English. This approach also takes the negations and intensity words into consideration. Cambria et al., (2012) develop a framework to create Chinese common sense knowledge for Chinese Sentiment Analysis. First, they merged different resources, i.e., ConceptNet, to produce a larger English knowledge base by using the blending technique. Second, they employed the translation tool to translate the whole knowledge base into Chinese version in terms of the fuzzy syntax-based model.

Although the Chinese sentiment analysis has been gotten attention for a decade (Ye et al., 2005, Ku et al., 2006, Wan 2008, Zhang et al., 2009, Liu 2012), to the best of our knowledge, there is no research investigating the sentiment of the topic-independent Chinese online news. Moreover, the news articles always have new events with the newest terms which created from the Chinese users, and the variety of the usage of the words associated with the newest terms makes the Chinese sentiment identification more difficult. Investigating the common features for identifying the opinion of the
Related Content

Neurofeedback: Refining the Methodology of Brain-Computer Interface Training
[www.igi-global.com/chapter/neurofeedback-refining-methodology-brain-computer/69459?camid=4v1a](www.igi-global.com/chapter/neurofeedback-refining-methodology-brain-computer/69459?camid=4v1a)

Morphological Contour Based Blood Vessel Segmentation in Retinal Images Using Otsu Thresholding
[www.igi-global.com/article/morphological-contour-based-blood-vessel-segmentation-in-retinal-images-using-otsu-thresholding/214896?camid=4v1a](www.igi-global.com/article/morphological-contour-based-blood-vessel-segmentation-in-retinal-images-using-otsu-thresholding/214896?camid=4v1a)
Timely Autonomic Adaptation of Publish/Subscribe Middleware in Dynamic Environments
www.igi-global.com/chapter/timely-autonomic-adaptation-publish-subscribe/68950?camid=4v1a

Model-Driven Approaches to Service Composition
www.igi-global.com/article/model-driven-approaches-to-service-composition/98597?camid=4v1a