DOI: 10.20472/IAC.2016.027.019

CHIHLI HUNG
Chung Yuan Christian University, Taiwan

YOU-XIN CAO
Chung Yuan Christian University, Taiwan

SENTIMENT CLASSIFICATION FROM WORD OF MOUTH DOCUMENTS BASED ON CHINESE COLLOCATIONS

Abstract:

Word of mouth (WOM) has become the main information resource while making business or buying strategies. Most WOM mining research studies focus on classification of WOM documents according to their sentimental orientations, i.e. positive and negative. Generally speaking a well-defined sentiment lexicon is used to provide the sentiment score for words. As a word may have different meanings when used in different domains so it may have different sentiment score. However such a lexicon is static and does not adapt to different domains. In this paper, we first build an adaptive Chinese sentiment lexicon from a real product review website. Then we identify feature words and opinion words of each sentence via the technique of mutual correlations between words. Based on association rules and mutual information, we extract the feature words and their associated collocation words. Finally the term frequency-inverse class frequency (TF-ICF) is used to extract word sentiment scores. According to experimental results, the usage and distribution of words are varied from different domains and our approach has a potential for Chinese WOM classification.

Keywords:

Word of Mouth; Sentiment Analysis; Opinion Mining; Association Rule; Sentiment Lexicon

JEL Classification: D80

Introduction

With popularization of the Internet, people are getting used to expressing their opinions, as well as the service or experience of products on the websites. This kind of electronic text is called word of mouth (WOM). Although WOM is informally expressed by consumers, it will affect the willingness of consumers to buy products, and when customers are going to make their strategic decision, WOM plays a key role as a marketing tool than any other promotion tools does (Herr, Kardes, Kim, 1991; Gilly et al, 1998). In the age of big data, people still take a long time to get the needed information, and it is also time consuming to digest the collected information by text reading. The text mining for sentiment analysis can explore and extract the subjective opinions, preferences, attitudes, etc, from WOM. On the other hand, the extracted and analyzed business information can be widely used as very important information source in many territories (Tang, Tan, Cheng, 2009).

To analyze and classify the non-structural WOM documents, a well-organized and wide scaled sentiment lexicon is usually helpful. SentiWordNet (Esuli, Sebastiani, 2006; Baccianella, Esuli, Sebastiani, 2010) and SenticNet (Cambria, Havasi, Hussain, 2012; Cambria et al., 2014; Cambria et al., 2016) are such lexicons. However, the same word in different sentences or domains may have different sentiment scores or even different sentiment orientations. Those predefined sentiment lexicons, such as SentiWordNet and SenticNet, are static, which may not handle user generated content on the Internet very well. On the other hand, the traditional WOM mining models suffer from the curse of dimensionality when the famous vector space model (VSM) in the field of text mining is used. Different parts of speech usually contain some relationships, for example, *cheap* (adjective) *price* (noun), *comfortable* (adjective) *room* (noun), etc. Unlike the traditional WOM mining model, which focuses on word only, the collocations are used in this research. Thus, we propose a collocation-based adaptive Chinese sentiment lexicon for sentiment classification of WOM documents.

Approach

The approach that we propose is divided into two stages. The first stage is to build a Chinese adaptive sentiment lexicon. There are three steps in the stage. Firstly, we collect WOM documents from the Urcosme cosmetic website (www.urcosme.com). Secondly, we preprocess those collected WOM documents based on a traditional text mining technique. Thirdly, we create the token-concept and token-sentiment matrices to build our proposed Chinese adaptive sentiment lexicon. The second stage is to build our WOM document vector for sentiment classification. There are two steps in this stage. Firstly, we combine a feature word with a sentiment word to build a collocation. Secondly, we build WOM document vectors for sentiment classification.

We consider a sentence a basic preprocessing unit. We extract only nouns and adjectives from each unit. A noun is a feature word regarding some specific feature of a product or a

service. An adjective is a sentiment word, which represents the degree how the user likes or dislikes the feature it describes. Thus, we combine an adjective with a noun to be a collocation. As a feature may be associated with many sentiment words and vice versa, we use the technique of association rule (Agrawal et al., 1993) to pick up the most suitable collocation. The following points describe our approach briefly.

1. Collecting WOM documents

   We crawl 113,178 WOM documents from www.urcosme.com and randomly choose 10,000 documents to build our model. The urcosme website is a famous cosmetic forum in Taiwan. Each WOM document is ranked from one-heart to seven-heart, which indicates the user's opinion from negative to positive.

2. Preprocessing

   As we are dealing with Chinese WOM documents, the task of Chinese segmentation is required. Besides the removal of HTML tags and punctuations, we use the famous Chinese segmentation system, i.e. CKIP (Chinese Knowledge Information Processing), to segment Chinese sentence into Chinese words. We then only keep nouns and adjectives and remove rest words with other parts of speech.

3. Building an adaptive sentiment lexicon

   We build the adaptive sentiment lexicon based on a straight concept, which the frequency of a token that appears in different conceptual categories denotes the relationship between it and its associated conceptual categories (Hung, 2013). For example, our cosmetic dataset has been pre-classified into 10 categories. A token shown in one specific category implies the relationship between it and this category. Thus, as our dataset has been ranked by the 7 heart rating system, a token shown in an one-heart document implies the relationship between this token and its negative sentiment orientation and vice versa. Therefore, we build two matrices to collect the relationship between tokens, categories and sentiments. These two matrices have also been used by Hung (2013).

4. Building a collocation

   The traditional vector space model (VSM) is used for the task of sentiment classification. However, this model usually suffers from the curse of dimensionality when dealing with a huge dataset. As different parts of speech play on different grammatical roles in a sentence. In terms of WOM, a noun is a subject or an object, which is a feature of a product or a service. An adjective contains much more sentiments than its associated subject or object. We combine these two parts of speech to form a collocation. For example, in terms of a noun, *price*, it might be *expensive*, *cheap*, etc. We try to use a collocation-based vector representation approach instead of a word-based vector representation approach for VSM. As a feature may associate with many sentiment

words and a sentiment word may associate with many feature words, we apply the association rule (Agrawal et al., 1993) to decide the most suitable allocation.

Association rule is an approach for discovering simultaneous relationships between two products. In other words, products *A* and *B* are bought at the same time. For the WOM sentiment classification task, a feature word *X* and a sentiment word *Y* are simultaneously used in a WOM. To select the most suitable allocation, the minimum thresholds on support and confidence should be decided. Based on the algorithm of association rule, support (1) is a value of how frequently the collocation in the dataset and confidence (2) is a value of how often the rule has been found to be true.

$$\text{support} = \frac{freq(noun \cap adjective)}{freq(all\ token)} \quad (1)$$

$$\text{confidence} = \frac{freq(noun \cap adjective)}{freq(noun)} \quad (2)$$

Experiments

In this research, except the 10,000 documents for building our model, we randomly choose 2,000 from the original 113,178 WOM documents to evaluate our proposed model. Table 1 describes the distribution of the 2,000 documents according to the 7-heart rating system.

Table 1: Sentiment distribution of 2,000 documents

Category	1-heart	2-heart	3-heart	4-heart	5-heart	6-heart	7-heart
Number of documents	27	111	107	287	474	680	314
Percent	1.35%	5.55%	5.35%	14.35%	23.70%	34.00%	15.70%

Source: Own Data

We use support vector machine from Weka 3.6 (www.cs.waikato.ac.nz/ml/weka) and classification results are evaluated by the classification accuracy criterion (3).

$$\text{accuracy} = \frac{the\ number\ of\ documents\ is\ correctly\ classified}{the\ number\ of\ total\ documents} \quad (3)$$

Based on our experiments, the accuracy of the traditional model, which is without using collocations, is 28.05%. Our proposed collocation-based model achieves an accuracy of 31.10%, which makes a slight improvement.

Conclusion

We propose a collocation-based approach for WOM sentiment classification. As this research is in its infant stage, both traditional model and our proposed collocation-based approach do not have outstanding results. According to our initial experiments, our proposed model has a slight improvement on classification accuracy for the task of WOM sentiment classification.

Acknowledgments

This work was partially supported by Ministry of Science and Technology of Taiwan, R.O.C., No. MOST 104-2410-H-033-039-MY2.

Reference

Agrawal, R.; Imieliński, T. and Swami, A. (1993) Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*. 207-216.

Baccianella, S.; Esuli, A. and Sebastiani, F. (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of International Conference on Language Resources and Evaluation*. 2200-2204.

Cambria, E.; Havasi, C. and Hussain, A. (2012) SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. *Proceedings of 25th International Florida Artificial Intelligence Research Society Conference*. 202-207.

Cambria, E.; Olsher, D. and Rajagopal, D. (2014) SenticNet 3: A Common and Common-Sense Knowledge Base for Cognition-Driven Sentiment Analysis. *Proceedings of Eighth AAAI Conference on Artificial Intelligence*. 1515-1521.

Cambria, E.; Poria, S.; Bajpai, R. and Schuller, B. (2016) SenticNet 4: A Semantic Resource for Sentiment Analysis Based on Conceptual Primitives. *Proceedings of COLING 2016*.

Esuli, A. and Sebastiani, F. (2006) SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of LREC 2006*. 417-422.

Gilly, M.C.; Graham, J.L.; Wolfinbarger, M.F. and Yale, L.J. (1998) A Dyadic Study of Interpersonal Information Search. *Journal of Academy of Marketing Science*. Vol. 26, No. 2, 83-100.

Herr, P.M.; Kardes, F.R. and Kim, J. (1991) Effects of Word-of-Mouth and Product-Attribute Information on Persuasion: An Accessibility-Diagnosticity Perspective. *Journal of Consumer Research*. Vol. 17, No. 3, 454-462.

Hung, C. (2013) Using Adaptive Contextual Sentiment Lexicons to Improve Quality Classification for Word of Mouth. *Closure Technical Report*, National Science Council of Taiwan, NSC 102-2410-H-033-033-MY2.

Tang, H.; Tan, S. and Cheng, X. (2009) A Survey on Sentiment Detection of Reviews. *Expert Systems with Applications*. Vol. 36, No. 7, 10760-10773.