Exploring Topic and Sentiment of Hotel Reviews

Ling Shen

Computer Science School, Wuhan Donghu University, Wuhan, China.

aleenapple@163.com

Keywords: Topic, Sentiment Lexicon, Hotel Review, Sentiment Strength

Abstract. Online review can help people getting more information about products. It offers a unique proposition for sentiment analysis. In this paper, a method for automatic extracting topic and it's sentiment of hotel reviews is proposed. Our method focus on a special case which means some explicit topics is already given. The unknown topics are extracted by means of given topics. Previous studies mainly focus on the polarity of the sentiment. Furthermore, in this paper, the sentiment strength is calculated by lexicon-based approach. Empirical experiments on hotel reviews demonstrate the advantage of the proposed method.

I Introduction

With the recent proliferation of Web2.0 sites and the rapid growth of user-generated-content on the internet, more and more internet users now publish online reviews to express their opinions and sentiments. Meanwhile, organizations are carrying out sentiment analysis and opinion mining of online reviews for decision making. Consequently, sentiment analysis has become a popular topic of many researchers.

It also makes it difficult for the seller to keep track and to understand customer sentiment. Previous studies mainly focus on the sentiment polarity mining [1] [2]. The sentiment polarity of the reviews is summarized. However, this is course-grained study.

Fig.1 Sample of Hotel Review

For instance, Fig.1 is a sample of hotel review from Tripadvisor.com. The website only list 5 topics. But the reviewer has more interest about food and WIFI other than those 5 topics.

On the whole, there are three contributions of our works:

1 A hotel-based lexicon collecting sentiment words is presented together with general sentiment lexicon.

2 A bootstrap algorithm is proposed to extract implicit topics in the reviews through explicit topics.

3 The Sentiment strength of each topic is calculated by linear regression model.

To our knowledge, this is first study exploring sentiment of implicit topics. The remainder of the paper is organized as follows. In Section 2, we first build a hotel-based sentiment lexicon, together with general sentiment lexicon to improve the accuracy; then we propose the algorithm to extract the implicit topics; and we present a method to calculate the sentiment strength of the sentiment sentence as well. In section 3 we present the experiments to demonstrate the efficiency of proposed methods. Finally, conclusions and directions for future work are given in section 4.

© 2015. The authors - Published by Atlantis Press

II The Proposed Method

Our goal in this paper is to summarize sentiment of the hotel reviews. To achieve this aim, 3 tasks should be considered. The first task is to build a sentiment lexicon. The second task is extracting the topics discussed in the hotel reviews. The third task is to determine the sentiment strength. We proposed different algorithms in the tasks mentioned above.

A Building Sentiment Lexicon

In the opinion mining and sentiment analysis research field, there are many sentiment lexicons such as SentiWordNet and MPQA [4] [5]. We choose it as general sentiment lexicon respectively. However, hotel reviews have their characteristics in natural language text. To improve the accuracy, sentiment lexicon special for hotel should be prepared in this task.

We extract the reviews from TripAdvisor.com from May 20, 2012 to June 20, 2012. 1500 reviews with four stars to five stars are regarded as positive samples while 500 reviews with one star to two stars as negative samples. Information Gain and CHI are popular methods in feature selection. We choose Information Gain to select the feature.

$$entropy(D) = -\sum_{j=1}^{|c|} P(c_j)\log_2 P(c_j), \sum_{j=1}^{|c|} P(c_j) = 1 \quad (1)$$

$$IG(t) = entropy(D) - entropy(c \mid t) \quad (2)$$

The Information Gain of the sentiment words in the hotel reviews are calculated after preprocessing. Then the sentiment words are listed in descending order respectively. We choose 1200 positive words and 460 negative words special for hotel manually. The whole lexicon is built by SentiWordNet plus sentiment word special for hotel and MPQA plus sentiment word special for hotel respectively. In the following experiment, we will compare the two lexicons.

Algorithm: Implicit Topic Extraction Algorithm
Input: A collection of hotel reviews $\{d_1, d_2 \dots d_{|D|}\}$, subset of explicit topic ET $\{T_1, T_2 \dots T_k\}$, vocabulary V, selection threshold p and iteration step limit I
Output: Set of whole topic T $\{T_1 \dots T_m\}$ $m \geqslant k$
Step 0: Split the reviews into sentences, $X = \{x_1, x_2 \dots x_n\}$, build a sentiment word list L
Step 1: Check each sentence in X; add sentiment words of topic T to L
Step 2: Match the sentiment word SW in each sentence X, record the match number Count (i)
Step 3: Assign the sentence a topic label by $a_i = \text{argmax}_i \text{Count}(i)$
Step 4: Calculate the IG measure of each word in V
Step 5: Rank the words under each topics with respect to their IG value and join the top p words for each topic into their corresponding topic keyword list L (T_i)
Step 6: If the topic keyword list is unchanged or iteration exceeds I, go to Step 7, else go to Step1
Step 7: Output the topic T $\{T_1 \dots T_m\}$ and its corresponding keyword list L (T_i)

Fig.2 Implicit Topic Extraction (ITE) Algorithm

B Extracting Implicit Topics

Although some topics are listed explicitly in the hotel review websites, there are also many important topics ignored by the webpage. For example, the review website may list room, price, location, sleep quality and service. However, some business customer care more about the WIFI, others care more about the food. All their opinions about these topics are expressed in the hotel reviews. In the previous works, these topics are ignored. They usually extract all the topics without using given topics.

In this section, we focus on the task of extracting implicit topics. We take advantage of the keyword derived from explicit topics to extract the implicit topics. Another method to extract the topics is topic model. Instead of treating each document as “a-bag-of-words” as in many models dealing with text documents, topic modeling assumes that a document is “a-bag-of-topics”, and the aim of topic modeling is to group each term in each document into a proper topic. A variety of probabilistic topic models have been proposed and LDA is one of the most popular topic modeling

methods [5]. LDA is unsupervised method, which requires no manually constructed training data. LDA's input is a matrix, and it outputs the document-topic distribution and topic-word distribution. Here document means review. In order to obtain the distributions and, two main algorithms were proposed. In the original paper of LDA [5], EM Algorithm was employed to solve the model. Gibbs Sampling [7] was proposed to estimate the parameters of the model later.

However in hotel review topic mining, some topics are already given. It is more convenient to extract the implicit topics based on the explicit topics. In comparison, the LDA is more complex. On the other hand, the hotel reviews mainly focus on about ten topics. The number of topics is very small. Since the number of the topics is a prior parameter, the comparative experiment shows that LDA is not efficient in mining the hotel review topics.

Here, we propose a bootstrap algorithm named ITE to extract the implicit topics through explicit topics. The method first collects sentiment keywords through explicit topics. Then rank the implicit topics using these sentiment keywords and join it to the topics.

The calculation of the IG also takes advantage of equations (1) and (2). The implicit topics and its corresponding sentiment words are extracted through ITE algorithm.

C Computing the Sentiment Strength

In last section, we already extract whole topics and corresponding sentiment words simultaneously. The third task is to determine the sentiment strength of each topic. In this section, we represent the review rating with a linear regression model of the topics. The strength of the sentiment word is also influenced by the distance between it and the topic [6]. We propose an equation computing topic score considering the negation words. All the scores are summed up using the following score function:

$$score(t) = \sum_{w_j \in t \wedge w_j \in V} (-1)^{c_N} w_j.SP \tag{3}$$

In equation (3), $w_j.SP$ means the sentiment polarity of the word w_j. A positive word is assigned the sentiment polarity score of +1, and a negative word is assigned the sentiment polarity score of -1. c_N means the number of negation word in one topic. While there is no negation, c_N equals 0. While there is one negation word, the polarity of the sentiment is reversed through multiply by -1. The rating of each review is a linear combination of different topics involved in it.

$$rating'(d) = \sum_{t_i \in d} score(t_i) * weight(t_i) \tag{4}$$

$weight(t_i)$ means the weight of topic t_i. For example, topics such as value, service, facility, WIFI, and food are mentioned in the review of Figure 1 in section 1. These topics have different weight, and the weight is unknown. If the weight of each topic is calculated, we will be aware of which topic the customers pay more attention to. We combine the equation (3) and equation (4) to get the equation (5) listed below.

$$rating'(d) = \sum_{t_i \in d} weight(t_i) \sum_{w_j \in t_i \wedge w_j \in V} (-1)^{c_N} w_j.SP \tag{5}$$

III Experimental results

In this section, experiments are presented to demonstrate the efficiency of the proposed methods. The hotel reviews are extracted from TripAdvisor.com from May 20, 2012 to June 20, 2012. We choose 2400 reviews randomly which include all the reviews mentioned in II. The stopwords are erased in the dataset. Then we use Porter Stemming tool to stem all the word.

A Sentiment Lexicon Comparison

Table 1. Classification Accuracy of Different Lexicon

Lexicon	Number of Lexicon	Classification Accuracy
SentiWordNet	6810	59.4%
MPQA	6400	60.5%
SentiWordNet + Hotel	6960	61.9%
MPQA + Hotel	6520	63.2%

B Implicit Topic Extraction Result

Table 2. Classification Accuracy with Different Topic Number

Number of Topic	Classification Accuracy	
	LDA	*ITE*
6	61.6%	64%
8	61.8%	65.8%
10	64.9%	67.2%
12	64.3%	68.5%
14	63.5%	64.6%
16	62.1%	-
18	61.7%	-
20	60.4	-

IV Conclusions

In this paper, the topic and corresponding sentiment of hotel reviews have been explored. First, the sentiment lexicon has been built; then the implicit topics have been extracted by means of explicit topics; in the end, the sentiment score has been calculated. Experimental results demonstrate the efficiency of proposed method.

Acknowledgment

This work is supported by Youth Fund of Wuhan Donghu University.

References

[1] Hu, Minqing, and Bing Liu. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD ACM, 2004.

[2] Jindal, Nitin, and Bing Liu. Opinion spam and analysis. Proceedings of the international conference on Web search and web data mining. ACM, 2008.

[3] Esuli, Andrea, and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. Proceedings of LREC. Vol. 6. 2006.

[4] Wilson, Theresa, et al. OpinionFinder: A system for subjectivity analysis. Proceedings of HLT/EMNLP on Interactive Demonstrations. Association for Computational Linguistics, 2005.

[5] Blei D, NG A Y, and Jordan M I: Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003.3(3):993-1022

[6] Ding, Xiaowen, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. Proceedings of the international conference on Web search and web data mining. ACM, 2008.

[7] Griffiths T, Stevyers M: Finding Scientific Topics. Proceedings of the Natural Academy of Sciences 2004.101:5528-5535