I.J. Education and Management Engineering, 2012, 6, 15-22
Published Online July 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2012.06.03

Available online at http://www.mecs-press.net/ijeme

A Research on Opinion Analysis for Book Reviews

[a]Na Zhai, [b]Fang Yuan, [c]Yu Wang

College of Mathematics and Computer Science, Hebei University Baoding, Hebei, China

Abstract

Product reviews are not only useful for trade companies to improve the quality of products, but also helpful for customers to purchase products reasonably, thus product reviews mining is valuable in application and research. In this paper, we devote the research on book reviews. We first propose a polarity dictionary construction method based on the improved CHI, and realizes dynamic addition of the dictionary; Second, the polarity calculation formula of the transitional complex sentences is improved to be applicable to book reviews. Considering that some book reviews have titles and these titles generally express the reviewers' opinion tendency, so we further propose an opinion polarity analysis method based on the titles and the improved polarity calculation formula of the heavy transitional sentences. The experimental results show that the approach proposed in this paper is effective.

Index Terms: dynamic dictionary; title polarity; polarity analysis; heavy transitional sentence

1. Introduction

With the rapid development of the network, online shopping has gradually penetrated into the daily life of people. In order to understand the users' evaluations of the products and services, trade companies provide a platform to users for airing their opinions. More and more users will publish their reviews on the web after shopping online [1]. As these reviews contain a lot of product information and the consumers' evaluations to products, so more and more potential consumers begin to understand the information of the required product by inquiring product reviews to decide whether to purchase products or not. Nevertheless, as the number of online reviews grows rapidly, there may be hundreds of or even more reviews for one product. It is difficult for consumers to rapidly get effective information. Therefore, it needs an effective method to analyze and summarize these reviews so as to make it easy for trade companies and customers to use these reviews.

2. Related Works

In 2004, Reference [2] first proposed a method to extract product features using association rule mining algorithms. Reference [3] carried out a more systematic research on product reviews, which not only mined

Corresponding author:
E-mail address: [a] zhaina19841201@sina.com, [b] yuanfang@hbu.cn, [c] wy@hbu.cn

product features, but also summarized the opinions. Opinion Observer system is built based on the pioneer work, which used the visualization method to analyze and compare products reviews [4].

Reference [5] obtained five dependency relations between the features and opinion words by syntactic parsing to generate feature-opinion candidates, and then transfered feature-opinion pairs mining into a classification problem.

At present, there are many researches on opinion mining based on the Chinese domain in domestic. Reference [6] mined product features. They not only extracted frequent features, but also used patterns to match and discover infrequent features. Reference [7] proposed the method which used collocations of words to extract the features and opinion words.

The above methods just focused on the features and opinion words in the reviews polarity analysis. They did not consider the influence of transition on the sentences polarity. Reference [8] pointed out that the transitional complex sentences and sentence groups may affect the whole ploarity of the texts, so he introduced transitional complex sentences recognization into the sentences polarity analysis.

Besides, the polarity dictionary constructed by this methods mentioned above is to be composed of fixed words. Thus it may cost a plenty of time to predict the polarity of each new word. In addition, they focused on the contents of reviews in polarity analysis, so it is easy to generate error analysis if this idea is used to analyze the whole reviews.

Based on the above considerations, we first use the improved CHI to construct the polarity dictionary, and realize dynamic addition of the dictionary; second, the influence of transition on the reviews polarity is considered in the polarity analysis, and we improve the polarity computational formula of the transitional complex sentence; finally, we consider the polarity of titles as the polarity mark of reviews for analyzing the reviews polarity.

3. Ploarity Analysis and Opinion Summarization of Book Reviews

3. 1 The Improved CHI-based Dynamic Polarity Dictionary Construction

1) The Improved CHI-based Dictionary Construction

CHI [9] is a common method of text feature selection, and its main idea is that words and categories are considered to accord with χ^2 distribution, and the χ^2 of words measures the contribution of the word to a category. The χ^2 computational formula of word i and category j is as follows [9]:

$$\chi_{ij}^2 = \frac{n \times (n_1 \times n_4 - n_2 \times n_3)^2}{(n_1 + n_2) \times (n_3 + n_4) \times (n_1 + n_3) \times (n_2 + n_4)} \qquad (1)$$

Formula (1) only denotes the correlation strength between the word and the category, but does not consider the contributions. If $n_1 \times n_4 - n_2 \times n_3 > 0$, then the contribution is positive, if $n_1 \times n_4 - n_2 \times n_3 < 0$, then the contribution is negative, so the improved χ^2 [10] is defined as:

$$\chi_{ij}^2 = sign(n_1 \times n_4 - n_2 \times n_3) \times \frac{n \times (n_1 \times n_4 - n_2 \times n_3)^2}{(n_1 + n_2) \times (n_3 + n_4) \times (n_1 + n_3) \times (n_2 + n_4)} \qquad (2)$$

Although the positive sign and negative sign of the original CHI is only reserved, whether the contribution of one word to one category is positive or negative can be distinguished.

Based on the ideas mentioned above, the CHI-DC algorithm of dictionary construction is as follows.

Input: The data belong to positive and negative categories after Word processing

Output: Positive and negative polarity dictionary

Algorithm steps:

1. Remove the stop words from each category.
2. for each word, count the document frequencies of the word occurring in each category.
3. Count the document frequencies of all words except this word occurring in each category.
4. Compute the χ^2 of the word using (2).
5. end.
6. For the words with positive contribution, the words which meet the threshold are added into the initial dictionary and remove the meaningless words to get the final dictionary.

2) Dictionary Dynamic Addition

Reviewers usually use many clauses with the same polarity to express the same opinion, and the words with the same polarity usually appear together. According to this characteristic, we extract words which express the reviewers' attitudes but have not been included in the dictionary. To extract unknown polar words, we need to get the co-occurrence relationship between the unknown polar words and known polar words. The co-occurrence factor s is defined as (3).

$$s = \frac{f_1}{f_2} \tag{3}$$

Where f_1 is the co-occurrence frequencies of an unknown polar word and positive words, f_2 is the co-occurrence frequencies of an unknown polar word and negative words. If this word often co-occurs with positive words, s is bigger, if $s > \sigma_1$, this word will be added into the positive dictionary; if this word often co-occurs with negative words, s is smaller, if $s < \sigma_2$, this word will be added into the negative dictionary; if the co-occurrence frequency of this word and the positive words approximates the co-occurrence frequency of this word and the negative words, s approximates 1, and this word is considered as a nonpolar word. As the denominator cannot be zero and it shows that this word only co-occurs with positive words when the denominator is zero, so we stipulate that s is equal to f_1 when f_2 is zero.

The DDA algorithm for dynamic addition of dictionary is as follows.

Input: The reviews after Word processing

Output: The words needed to add

Algorithm steps:

1. for each review, determine the polarity of each opinion clause via the known polar words and record the locations of polar words.
2. for each adjective without containing in the polarity dictionary.
3. if this word is located in the positive or negative clauses.
4. if this word is located among the positive words and not modified by a negation word, or located among the negative words and modified by a negation word.
5. this word co-occur with positive words.
6. else this word co-occur with negative words.
7. else if this word is located between the positive clause and negative clause.
8. if there is a transition in the sentence, identify the relationship between this word and the opinion sentences before transition and after transition.
9. if this word is in the positive clause before transition or after transition and is not modified by a negation word, or it is in the negative clause before transition or after transition and is modified by a negation word.
10. this word co-occur with positive words.
11. else this word co-occur with negative words.
12. search in the unknown polar word lists whose words co-occur with positive words or negative words.

13. if this word is in the lists, then the co-occurrence frequency will be adjusted.
14. else this word will be added into the corresponding list.
15. end.
16. Obtain the co-occurrence frequency of each word from unknown polar word lists. If $f_2 = 0$, then $s = f_1$, else get the s of the word by (3), and then add the words which meet the threshold into the corresponding dictionary.

In steps 12-14, in order to enhance the speed of searching in the lists, we set the upper limit T and the threshold σ_3. When the size of any unknown polar word list reaches T, we will remove the word whose s meets $|s-1|<\sigma_3$ from the two unknown polar word lists to reduce the size of the lists.

3. 2 Reviews Polarity Analysis Based on the Titles and the Improved Polarity Computational Formula of the Heavy Transitional Sentences

Current methods all focused on the contents of reviews when analyzing the polarity of reviews, and they only classified the review sentences. However, it is easy to generate the wrong polarity identification when this method is applied to analyze the whole reviews.

We discover that some book reviews have titles which usually indicate the attitude of the whole reviews. Thus, we use the title to analyze the polarity of reviews.

Besides the problem mentioned above, users usually use the transitional complex sentence in the title and content of reviews to express their opinions on books. If we simply determine the polarity according to opinion words, then the wrong polarity identification will be made for the heavy transitional sentence.

In order to solve the problem mentioned above, we introduce transitional complex sentences processing into the polarity calculation of book reviews. Reference [8] gave the formula used in determining the transition complex sentence polarity:

$$so = \alpha \times SO(SubSt'_{pos}) + SO(SubSt_{pos}) \tag{4}$$

But it only weakens one clause before transition. If the polarity value of the clause which has not been weakened is greater than the clause after transition, then the wrong polarity identification will be caused because the reviewers mainly want to express the opinion after transition. Similarly, the formula only emphasizes one clause after transition. In Addition, if the clause before transition does not have polarity, then it also cannot emphasize the opinion after transition.

Based on the above considerations, (4) is not suitable to be applied to book reviews in computing the polarity value of the transitional complex sentences. Therefore, (4) is improved in this paper. The improved polarity computational formula of the heavy transitional sentence is as follows:

$$S_p = w_1 \times \sum_{i=1}^{n_1} Subs_i + w_2 \times \sum_{j=n_2}^{n} Subs_j \tag{5}$$

Wher $Subs_i$ and $Subs_j$ is the polarity value of clause i before transition and the polarity value of clause j after transition, respectively, w_1 and w_2 denote the weights of sentence before transition and after transition in determining sentence polarity, respectively.

To sum up, we propose a review polarity analysis approach based on the titles and the improved polarity computational formula of the heavy transitional sentences.

The RPA algorithm of review polarity analysis as follows.

Input: The titles and contents of reviews after Word processing.

Output: The pros, cons or neutral set.

Algorithm steps:

```
1. for the title of each review, the polarity value of title sum=0.
2. for each opinion word p in the title, compute its contextual polarity.
3. determine the polarity of each clause in title according to the polarity of p.
4. if there is heavy transition sentence in title.
5. use (5) to get the polarity value of the title.
6. else accumulative the polarity value of each clause to get the polarity value of the title.
7. if (sum>0), divide the review into the pros set.
8. else if (sum<0), divide the review into the cons set.
9. else if (sum=0), further analyze the review content and the polarity value of review Csum=0.
10. { for each opinion sentence,  polarity value Sp=0.
11. for each opinion word p in the opinion sentence, compute its contextual polarity.
12. determine the polarity value of each clause according to the polarity value of p.
13. if the opinion sentence is a heavy transition sentence.
14. use (5) to obtain the Sp.
15. else accumulate the polarity value of each clause to obtain the Sp.
16. obtain the Csum by accumulating the Sp of every opinion sentences.
17. divide the review into the pros, cons or neutral set according to the Csum.}
18. end.
```

4. Experiments

4. 1 Data

The experimental data are from http://www.amazon.cn and http://www.dangdang.com, and we randomly select 761 review titles and 720 reviews from the computer book reviews.

4. 2 Polarity Dictionary Construction

We use the method proposed in this paper to construct dictionary and realize dynamic addition. The experimental results are presented in Table I.

Table I. Experimental results of polarity dictionary construction

Polarity	***Artificial recognition***	***Algorithm recognition***	***Correct recognition***	***Prec***	***Rec***
Positive	76	71	69	97.18%	90.79%
Negative	61	57	54	94.74%	88.52%

The results show that the proposed method can effectively divide the polarity words. However, in whole the recall of positive words and negative words are not very high, because some words that only appeared once in the corpus cannot be identified, but they could be identified with the increasing of experimental data.

For the dictionary constructed above, we dynamically add the new words to this dictionary by using the method proposed in the paper. We add the words “conscientious” and “comfortable” into the positive dictionary, but no words for negative dictionary.

4. 3 Predicting the Polarity of Reviews

We use the proposed method to analyze the polarity of reviews. To verify the reasonability and effectiveness of our method, we also use three other methods to compare. The results are listed in Tables II, III and IV, respectively.

Tables II. Results of polarity analysis of reviews without considering titles and heavy transitional sentences

		Origi-nal polarity	***Algorithm recogn-ition***	***Correct recogn-ition***	***Prec***	***Rec***
Ama-zon	Positive	200	200	177	88.50%	88.50%
	Negative	150	131	117	89.31%	78.00%
Dang-dang	Positive	225	234	201	85.90%	89.33%
	Negtive	145	118	103	87.28%	71.03%

Table II shows that the precision and recall are not very high. Because it is likely to use transitional sentences and introduce comparisons between the products when they commented products. This is easy to generate the wrong polarity identification when directly analyzing the reviews.

Table III. Results of polarity analysis of reviews without considering titles but considering heavy transitional sentences

		Origi-nal polarity	***Algorithm recogn-ition***	***Correct recogn-ition***	***Prec***	***Rec***
Ama-zon	Positive	200	191	177	92.67%	88.50%
	Negative	150	139	125	89.93%	83.33%
Dang-dang	Positive	225	225	201	89.33%	89.33%
	Negtive	145	122	108	88.52%	74.48%

Table III shows that the precision and recall have been improved compared with Table II, because the method considers the heavy transitional sentences. This proves that it is necessary to consider heavy transitional sentences.

Table IV. Results of polarity analysis of reviews using the method that based on titles and the polarity computational formula of heavy transitional sentences [8]

		Origi-nal polarity	***Algorithm recogn-ition***	***Correct recogn-ition***	***Prec***	***Rec***
Ama-zon	Positive	200	205	187	91.22%	93.50%
	Negative	150	142	130	91.55%	86.67%
Dang-dang	Positive	225	220	209	95.00%	92.89%
	Negtive	145	148	133	89.86%	91.72%

Table IV shows that the precision and recall have obviously increased, because the method considers the polarity of titles. This proves that the impact of the titles on the determination of the polarity of reviews.

Table V. Results of polarity analysis of reviews based on the title and the improved polarity computational formula of the heavy transitional sentence

		Origi-nal polarity	***Algorithm recogn-ition***	***Correct recogn-ition***	***Prec***	***Rec***
Ama-zon	Positive	200	200	188	94.00%	94.00%
	Negative	150	145	135	93.10%	90.00%
Dang-dang	Positive	225	220	212	96.36%	94.22%
	Negtive	145	148	136	91.89%	93.79%

Table V shows that the precision and recall are very good when using the method proposed in this paper. Comparing with Table III, the results prove the effectiveness of the proposed approach. Comparing with Table IV, the improved formula in this paper can be more suitable for the book reviews than the one used in [8].

5. Conclusions

For book reviews, we first propose a method of constructing polarity dictionary, which is based on improved CHI, and realize dynamic addition of the dictionary, which solved the problem of fixed dictionary to some extent. Second, considering the problems caused by only analyzing the reviews in reviews polarity analysis and the impact of heavy transitional sentences on the polarity of reviews, we improve the polarity computational formula of heavy transitional sentence and propose a polarity analysis method based on the titles and the improved polarity computational formula of heavy transitional sentences. Finally, Experimental results indicate that the proposed method is effective.

Acknowledgements

This work was supported in part by the Key Research Plan of Education Office of Hebei Province (ZH200804).

References

[1] WU Xing, HE Zhong-shi and HUANG Yong-wen. Survey on Product Review Mining. Computer Engineering and Applications, vol. 44, pp. 37-41, 2008. (in Chinese)

[2] Hu M, and Liu B. Mining Opinion Features in Customer Review. To appear in AAAI'04, pp. 755-760, 2004.

[3] Hu M, and Liu B. Mining and Summarizing Customer Reviews[C]. KDD'04, pp. 168-177, 2004.

[4] Liu B, Hu M, Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th international conference on world wide web, Chiba, Japan, pp. 342-351, 2005.

[5] Gamgarn Somprasertsri, Pattarachi Lalitrojwong. Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. Journal of Universal Computer Science, vol. 16, pp. 938-955, 2010.

[6] Wei Wei, Hongyan Liu, Jun He, Hui Yang, Xiaoyong Du. Extracting Feature and Opinion Words Effectively from Chinese Product Reviews. International Conference on Fuzzy Systems and Knowledge Discovery, pp. 170-174, 2008.

[7] Yun-qing Xia, Rui-reng XU, Kam-fai Wong and Fang Zheng. The Unified Collocation Framework for Opinion Mining. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 2007.

[8] Xiao Wei. Semantic Based Sentiment Classification on Blog Community. Shanghai: Shanghai Jiaotong University, 2007. (in Chinese)

[9] Hinrich Schutze, David A.Hull and Jan O. Pedersen. A Comparison of Classifiers and Document Representations for the Routing Problem. In Proceedings of SIGIR.95, 18'h ACM International Conference on Research and Development in Information Retrieval, pp. 229-237, 1995.

[10] Yu Wang, Zheng-Ou Wang. Text Categorization Rule Extraction Based on Fuzzy Decision Tree. Machine Learning and Cybernetics, Proceedings of 2005 International Conference on, vol. 4, 2122-2127, 2005.