# The study on Web product reviews mining based on an improved text categorization algorithm

**Dongbin Hu[1] Lixia Luo[1] Lihua Xu[1]**

[1]School of Business, Central South University, Changsha 410083, P. R. China

## Abstract

With the rapid development of the Internet and e-business, more and more netizens tend to publish comments of the various products on the net. However, to obtain information from these data will face some problems: huge, emir-structured and non-structured data. In this regard, performing data mining is particularly important. This paper puts forward to a new method of classifying these reviews, and then introduces the mining process of the product comments based on the text classification algorithm; finally, verifying this new method's feasibility and accuracy with two products of IT168 website's comments. The example analysis results show that the method is practical, and also get good classification results.

**Keywords**: Web product reviews mining, Polarity, Polarity strength

## 1. Introduction

Online shopping has become a new fashion. Many shopping sites, such as E-commerce sites, shopping guide websites, offer consumers to perform comments on the platform, by this way consumers could timely give feedback to merchants and potential users. Because the data of comments is semi-structured, non-structured, and huge and disrupt distributing, the traditional technology of data mining cannot deal with these data [1]. Product review mining has become a research focus in recent years. Many text mining algorithms can be applied to product review mining, including the classification algorithm based on association rules [2], the classification algorithm based on support vector machine [3] and so on. But these algorithms have some shortcomings, such as the complexed process and the satisfacted results [4]. Therefore, this paper introduces an improved points device classification algorithm on automatic text categorization.

## 2. The improved points device classification algorithm

The main difference between this new algorithm and the traditional one is that the latter assigns points to each characteristic through studying the training dataset; however, the former, through polarity studies on classification characteristics [5], can get polarity strength, and then combining with HowNet term set of sentiment analysis, which is the most authoritative knowledge base. Besides, the improved one divides the text into three categories of positive, negative and neutral with scores; however, the traditional one has only two. It is worth emphasizing that scores indicated attitude strength is built by experience with characteristics a product. Moreover, in real work, we can do many changes, for example, we can first grade scores according to some criterion, and then divide the comments into more classes, and the core mechanic can not do any change. Specific formulas can be described in the following section.

## 3. Web product reviews mining steps

In this paper, Web product reviews mining have six procedures such as Reprocessing, Product features recognition, Polarity identification and Polarity strength determination, Polarity analysis of sentences, Classification and evaluation of results, Results display and application.

### 3.1. Preprocessing

Just like text mining, Web product review mining steps based on improved points device classification algorithm has four methods, including Sub-clause, Chinese segmentation, POS tagging and Removing stop words.

（1）Sub-clause

Because user usually use nonstandard punctuation, so we set some Symbols, such as ".", "?", "!"," ... ... "" ~ "and so on, as the boundary of the initial sentence[6].

（2）Chinese word segmentation

In this paper, we use ICTCLAS, developed by the Institute of Chinese Academy of Sciences computing systems (Download site: http://ictclas.org/index.html). Experiment results demonstrate that its precision reached the level of 97.85%, and its recall is above 90%. ICTCLAS system is the most advanced Chinese word segmentation system.

（3）POS (Part-of-speech ) tagging

POS tagging is to determine the grammatical category and parts of speech of each word in a given sentence, and then to marked theses words with a part-of-speech identifier, such as noun, verb, adjective and adverb.

（4）Stop-word filtering

We take feature words and words with high frequency or low frequency as stop words, which play a supporting role in the text. We make use of the method based on word-frequency to remove them, and also implement a "stop word list" instead [7].

### 3.2. Product features recognition

Features here are the words about product's features, functions, and parts. Feature recognition consists of two steps: feature word extraction and feature marking.

（1）Feature word extraction

The main task of feature word extraction is extracting feature words which present product characteristics from a large number of comments .Because most products feature words are nouns, so we select nouns for statistical analysis only. We count the frequency of each noun in all comments and select nouns that have higher term frequency as candidates for feature words. Then, we tease out rational feature words.

（2）Feature tagging

If a comment contains only a single feature word, name the sentence as simple feature sentence, and tag it with the feature word that it contains. If a comment contains more than 2 features of words, we call complex feature sentence, and then we can divide it into two or more clauses according to corresponding punctuation, so here, each clause is a simple feature sentence. However, sometimes user’s reviews may not be for a specific feature, but t he whole product, so that the reviews don’t have any feature words. In this case we define a “entirety” to tag this kind of reviews.

### 3.3. Polarity identification and Polarity strength determination

In general, we call words with emotional tendencies as polarity words, being divided into three categories: ameliorative word, neutral word and derogatory word. Just as we know, people can easily judge a word is ameliorative or not, but for computers, it is a very difficulty. Therefore, we usually combine manual with automated methods [8]. The process is a two step process.

（1）Establishment of three Thesaurus based on HowNet: Good (positive), Bad (negative), Normal (neutral).

HowNet (Download site: http://www.keenage.com/html/c_index.html) has classified words into positive, derogatory, and neutral [9]. These words' polarity need to be marked manually. In accordance with the general marking method, the polarity strength is divided into three levels: 1.0, 0.5, 0.0, and we take the word with 0.0 as polarity strength of neutral words.

（2）Judgment the polarity of a new word

According to SO-PMI [3], Pointwise Mutual Information is a useful Information Measure Method, and it refers to the correlation of two events set, the probability formula can be described as follows:

$$\text{PMI}（\text{word}_1, \text{word}_2）= \log_2\left(\frac{\text{P}（\text{word}_1 \,\&\, \text{word}_2）}{\text{P}（\text{word}_1）\text{P}（\text{word}_2）}\right) \quad (1)$$

There, $\text{P}（\text{word}_1）$ is the appearing probabilities of $\text{word}_1$, in the same way, $\text{P}（\text{word}_2）$ is the appearing probabilities of $\text{word}_2$ and $\text{P}（\text{word}_1 \& \text{word}_2）$ is the co-appearing probabilities of $\text{word}_1$ and $\text{word}_2$. So, we can get new word's polarity strength through calculate the Point-wise Mutual Information of new word and seed word, existing in the Thesaurus. The formula as follows:

$$Score(word) = \sum_{pword \in pset} PMI(word, pword) - \sum_{pword \in nset} PMI(word, pword) \quad (2)$$

There, $Score(word)$ means polarity strength of $word$ , $word$ expresses new word, $pword$ expresses seed word, $pset$ expresses the set of words from Good , and $nset$ expresses the set of words from Bad.

### 3.4. Polarity analysis of sentences

In this part, we first judge the polarity of participles, picked out from the given sentence. If polarity words are present, we can get its polarity and polarity strength. Second, we should detect whether privative and adverb of degree or not, if so, we do the same progress just like previous. Otherwise, we mark privative with a weight of -0.5 and mark adverbs of degree with one weight of 1.5，1.2，0.8，0.6 according their meaning. Finally, we set the average of and polarity strength of each word as the polarity outcome of the sentence, figured up according to the following expression:

$$polarityscore = \frac{\sum_{i=1}^{m} score(p_i)\ pt_i dt_i + \sum_{j=1}^{n} score(n_j)\ pt_j dt_j}{m+n} \quad (3)$$

There, $\text{score}\,(p_i)$ represents polar strength of the ith positive polarity word; $\text{score}\,（n_j）$ is negation sign. If there are privatives before polarities, $pt_i$ equals to -0.5; if not, $pt_i$ equals to 1.0. And $dt_i$ represents the weight of degree adverbs, if there is the degree adverbs, $dt_i$ equals to the weight of the degree adverb which modifies ith positive polarity word .

### 3.5. Classification and evaluation of results

In this paper, we use two-stage classifications: the first stage, we set product feature words as categorical attribute, and then divide the comments into the corresponding categories. The second stage, we set polarities as indexes, and then classify comments into three categories of positive, negative and neutral, adopting the improved points device classification algorithm.

$$class(m_i) = \begin{cases} c & polarityscore(m_i) > 0 \\ c' & polarityscore(m_i) < 0 \\ c'' & polarityscore(m_i) = 0 \end{cases} \tag{4}$$

There, $class(m_i)$ represents the classification level of the ith comment; $\mathrm{polarityscore}(m_i)$ represents the polar strength of the ith comment; $c$ means this comment belongs to the positive category; $c'$ is negative category; $c''$ is neutral category. Meanwhile, we can also calculate out the average polar score of each product feature according to the following formula:

$$\text{avgscore}（A）= \frac{\sum_{i=1}^{m} p_i + \sum_{j=1}^{n} q_j}{N} \tag{5}$$

There, $p_i$ represents the score of the ith positive comment of feature A, $q_j$ represents the score of the jth negative comment of feature A, m is the number of positive comments of feature A, n is the number of negative comments of feature A, N is the number of total comments of feature A. And then, we could figure out the average polar score of the total product features according to the following formula. To be emphasized, they have the same formula, but have different meanings.

$$\text{avgscore}（\text{entirety}）= \frac{\sum_{i=1}^{m} p_i + \sum_{j=1}^{n} q_j}{N} \tag{6}$$

There, $p_i$ represents the score of the ith positive comment, $q_j$ represents the score of the jth negative comment, m is the number of positive comments, n is the number of negative comments, N is the number of total comments.

In evaluation of classification results, we use the accuracy and the recall rate as evaluation indexes. First of all, we do some conventions:

- The number of testing documents that positive cases were correctly classified;
- The number of testing documents that negative cases were wrongly classified;
- The number of testing documents that positive cases were wrongly classified;
- The number of testing documents that negative cases were correctly classified.

The higher precision and recall rate, the better text classification effect. In general, if a classification accuracy reaches 80% or more, then that the method is feasible. And the calculation formula is given out as followings:

$$P = \frac{a}{a+b}, \quad R = \frac{a}{a+c} \tag{7}$$

There, P represents accuracy; R represents recall rate.

### 3.6. Results display and application

Mining results display in the form of tables and graphics, by this way, users can obtain information they need; In order to make results concise and easy to understand, they can also be divided into several categories to display. When applying, users can not only obtain polar category and polar strength of their concerned products, but also compare with different products.

## 4. The Application Analysis of Text-based Classification for Product Reviews Mining

### 4.1. The Background of Examples

The reviews data used in the paper come from the website: http://www. itl68. com. IT168 is a renowned brand on the area of IT purchase products in china, and is the one of the largest and most authoritative Shopping guide information websites. There are two products' comments downloaded from the IT168 website, Nokia 5320XM of 206 reviews and the Nokia 5800XM of 205 comments respectively. After preprocessing and feature tagging, we get 136 product features and 2369 simple feature sentences for Nokia 5320XM; at the same time, we get 128 product features and 2186 simple feature sentences for Nokia 5800XM.

### 4.2. The Mining results and analysis

Because of the large amount of data, this paper only show the mining results of the product features which have most amount of reviews. The mining results of the product feature of the Nokia 5320XM's battery and Nokia 5800XM's screen are shown in Table 1 and Table 2 respectively.

| Polarity | Positive | Negative | Neutral |
|---|---|---|---|
| The original polarity (piece) | 96 | 653 | 63 |
| Experimental polarity (piece) | 112 | 605 | 72 |
| Experimental error points (piece) | 21 | 86 | 13 |
| Accuracy rate (%) | 81.25 | 85.8 | 81.9 |
| Recall rate (%) | 94.8 | 79.5 | 93.7 |

Table 1: Battery of Nokia 5320XM.

| Polarity | Positive | Negative | Neutral |
|---|---|---|---|
| The original polarity (piece) | 749 | 125 | 43 |
| The experimental polarity (piece) | 712 | 102 | 39 |
| The Experimental error (piece) | 83 | 16 | 9 |
| Accuracy rate (%) | 88.3 | 84.3 | 76.9 |
| Recall rate (%) | 84 | 68.8 | 69.8 |

Table 2: Screen of Nokia 5800XM.

In the case of the product features are not considered, the option words for each product are divided into three categories, the result is shown in Table 3.

| Product Name | The average accuracy | The average recall |
|---|---|---|
| Nokia 5320XM | 85.6 | 82.9 |
| Nokia 5800XM | 83.1 | 81.2 |

Table 3: Review mining results of each product.

According to SVM [3], we obtain the classification of reviews of two products as follows:

| Polarity | Positive | Negative |
|---|---|---|
| The original polarity (piece) | 986 | 1383 |
| The Experimental polarity (piece) | 904 | 1185 |
| The Experimental error points (piece) | 165 | 203 |
| Accuracy rate (%) | 81.7 | 82.9 |
| Recall rate (%) | 74.9 | 71 |

Table 4: Nokia 5320XM of SVM.

| Polarity | Positive | Negative |
|---|---|---|
| The original polarity (piece) | 1685 | 501 |
| The Experimental polarity (piece) | 1468 | 423 |
| The Experimental error points (piece) | 265 | 84 |
| Accuracy rate (%) | 81.9 | 80.1 |
| Recall rate (%) | 71.4 | 67.7 |

Table 5: Nokia 5800XM of SVM.

By this method, Nokia 5320XM has the average accuracy of 82.4%, and the average recall of 72.6%. Meanwhile, Nokia 5800XM has the average accuracy of 81.5%, and the average recall of 70.5%. The results from the above show that the classification results of product reviews improve obviously and the precision attains to 80%. So the result of new method is superior to SVM. By the simple statistical methods, the average accuracy of the sentence semantic attains to only 70%; Form the experimental results, it prove that, in determining the sentence polarity, only using statistical methods is not reliable.

### 4.3. The Mining results and the Application Analysis

By calculating the average polarity of the comments of each product feature, we can get the evaluation of each product features. The polarity scores of three features of Nokia 5320XM and Nokia 5800XM are enumerated in the following. Three characteristics of the two products in table 6:

| **Product** | **Battery** | **Function** | **Price** |
|---|---|---|---|
| Nokia 5320XM | -2.3 | +1.2 | +1.1 |
| Nokia 5800XM | -2.9 | +1.8 | +0.7 |

Table 6: The polarity strength of Nokia 5320XM and Nokia 5800XM.

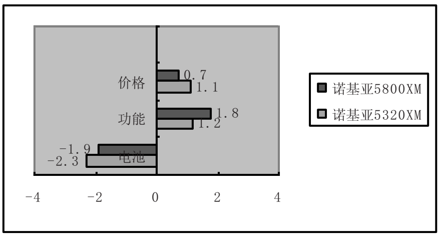

Fig.1: The comparison diagram of three features.

According to the above mining information, users can easily get the following information: Nokia 5800XM has a powerful function and reasonable price, but the battery is not very good; Nokia 5320XM has relatively complete function, but the battery is not very good and the price is more higher than the former; in the end, the same advantages of the two products is their price and function, but the battery rating is low, it is needed to improve. When making decisions, consumers can refer to these mining results to compare the two products' advantages and disadvantages. On the other hand, the mining information can also help manufacturers and businesses improve marketing strategies or production decisions.

## 5. Conclusion

The depth study of product review mining plays active roles in Strategic Decision-Making, Customer Relationship Management, and Enterprise Planning [10].This paper puts forward an improved point device classification algorithm, and then analyzes steps of web product reviews mining based on this method in details. Through this method the comments can be expressed as three polarities and the corresponding polarity strength. By this way, we can not only understand character condition of the given product, but also users' attitudes even the attitude strength towards the product. Finally, we use an example of IT168 to prove that this method is effective and feasibility. So, this paper can provide a reference for product reviews mining system.

## References

[1] Wu Xing, He Zhongshi, Huang Yongwen. Summary of the Research of product reviews mining [J]. *Computer Engineering and Applications,* pp. 37-41,2008,44（36）

[2] Wang Yuanzhen, Qian Tieyun, Feng Xiaonian. Asso-ciation rules based automatic Chinese Text Categorization [J]. *Mini-micro Systems,* 2005，26(8): 1380-1383.

[3] Huang Yongwen, He Zhongshi, Wu Xing. Categories access of users rviews[J]. *Computer Applications,* 2009,29(3):846-849.

[4] Zhuang Li，Jing Feng，Zhu Xiaoyan． Movie review mining and summarization[C]. *CIKM，* 2006： 43-50.

[5] Wang Chao, Lu Jie, Zhang Guangquan. A semantic classification approach for online product reviews[C].*Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*，French，2005.

[6] Christopher D, Manning H S. The statistical founda-tion of Natural Language Processing. Beijing: Electronic Industry Press, pp.67-121,2005.

[7] Liang Nanyuan. Written Chinese word segmentation system-CDWS[J]. *Journal of Chinese Information Processing*, 2007，32(2)：1-4.

[8] Kim SM，Hovy E. Identifying and analyzing judgment opinions[C].*Human Language Technology Con-ference of the North American Chapter of the Association of Computational Linguistics Proceedings*， USA，2006.

[9] Turney PD，LITTTMAN ML. Measuring praise and criticism：inference of semantic orientation from association[J]． *ACM Transactions on Information System*，2003，21(4)：315-346.

[10] Wang Yan, Zhang Fan. The design and implementa-\tion of information retrieval system based on web text mining[J], *Journal of the China Society for Scientific and Technical Information*,2007,26(3):339-343