Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MINING FREQUENT AND INFREQUENT FEATURES FROM CHINESE CUSTOMER REVIEWS

LI SHI, YU MING
School of information and computer engineering, Northeast Forestry University, Harbin 150080,
Heilongjiang, China

ABSTRACT

Customer reviews serve as a feedback mechanism that can help suppliers enhance their products and services, then gain competitive advantages. Mining Product features from reviews are expected to further investigate the views and attitudes of customers. This study is focus on one subtask of sentiment analysis. We want to extract the product frequent and infrequent features from Chinese customer reviews. Our approach is based on associated rule technique, and we further propose a algorithm which integrated the self- construct features datasets from websites to identify infrequent features. Experiments are conducted by using the reviews which download from Internet as corpus. Results proved that the algorithm will improve the performance of product features extraction, which will be helpful for identifying the real concern of customers.

Keywords: *Customer Reviews; Product Features; Data Mining*

1. INTRODUCTION

With the e-business arising in recent ten years, the reviews with a user-oriented content play an important role instead of the word-of-mouth in the offline world. According to Deloitte and Touche USA LLP's [1] data collected in a survey, more than 80% of those who read online reviews said that their purchasing decisions were directly influenced by the reviews. However, the number of reviews is growing rapidly and the contents are more complicated, as the interaction between the user and enterprise becomes more frequent and in-depth. It is very difficult for customers, merchants and manufacturers to retrieve valuable knowledge from customers' reviews. Some technologies are needed to enhance the accuracy and convenience of mining method. Mining valuable information from customers' reviews, which is review mining technology just aim at this problem [2~5]. In English reviews area, Researchers have made some successful progress but few studies have been conducted to Chinese customer reviews on the internet [6]. Because of the differences in characteristic between the two languages of Chinese and English, existing English oriented approaches were hard to imply directly on Chinese. This work just focus on Chinese customer reviews on the Internet, and explore the technology of product features mining, in order to provide a more convenient and scientific tool for enterprises and customers in Chinese e-commerce field.

In previous study, we present an unsupervised approach to extract product features from Chinese customer reviews [6]. It is based on the theory of association rules, in particular, Apriori algorithm to extract frequent itemsets as candidate product features. But the method has not been considered the infrequent features yet. However, some product features may not often appear in the reviews so that the candidate collection of features would not be complete. For example: mobile feature "software compatibility". As a smart phone, "software compatibility" is a very important feature. In contrast, this feature is professional, and for most users who use mobile phones may overlook it. As a result, it is not hot in the reviews and difficult to be extracted by the Apriori algorithm. This paper proposed a method to identify the infrequent features and promote the performance of feature extraction from Chinese reviews.

Table 1
Feature Information

Explicit Features	Examples
Properties	*ScannerSize*
Parts	*ScannerCover*
Features of Parts	*BatteryLife*
Related Concept	*ScannerImage*
Related Concept's Feature	*ScannerImageSize*

In the following section we describe previous work on the task of product feature and opinion extraction. People can express their views for anything, such as products, individuals,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

organizations and so on. Here we use a common object to represent the entity of the evaluation [7], so each object can be decomposed into hierarchical according to affiliation relationship. An entity e is a product, person, event, organization, or topic. The e is represented as a hierarchy of components, sub-components, and so on. Each node represents a component and is associated with a set of attributes of the component. Table 1 shows the entity examples of product features extracted from reviews [3].

This technique is the premise of sentiment classification attached to the specific features expressed by customer in the online reviews. Therefore, performance of mining result is important. Mining features from the English reviews has made some achievements. Similar to the development of sentiment analysis techniques, there are also two kinds of mining method, supervised [8] and unsupervised method [2, 6, and 9].

M. Hu and B. Liu focus on mining opinion and product features from customer reviews in 2004[2]. They use association rule and tested the method by internet reviews of some digital product and books, such as mobile phones, digital cameras, achieved 80% precision and 72% recall [2,9]. On this basis, they take a follow-up study to identify the polarity of opinion for the products features [7,9]. L. Zhang used B. Liu's feature-based opinion mining model to identify noun product features that imply opinions [10]. They found that in some domains nouns and noun phrases that indicate product features may also imply opinions. In many such cases, these nouns are not subjective but objective. Their involved sentences are also objective sentences and imply positive or negative opinions.

Some researchers have used other approaches to mining features. N. Kobayash adopted a semi-automatic method of extracting the cycle of product characteristics and user opinion, but needs a large number of manual efforts [11]. A.M. Popescu, who used KonwItAll systems which developed by O. Etzioni, compute the value of mutual information (PMI) of features and relations (discriminators). They use Bayesian classification to extract product features. The result of them contrast with Hu' performance is raising the precision rate (an average of 22%), but recall is decreased (an average decrease of 3%) [3]. In addition, Jian Liu focus on reviews in which customers compared variety of products [12]. Similar with A.M. Popescu's research, they adopt supervised method and made a training dataset first; it will need reading customer reviews one by one, which is contradict with purpose of automatically reviews mining. Qiu et al. proposed a method called double propagation that uses dependency relations to extract both opinion words and product features [13]. Ref [14] proposed opinion feature extraction based on sentiment patterns, which takes into account the structure characteristics of reviews for higher values of precision and recall.

As mentioned previously, in term of the big difference in language style and grammar between Chinese and English, the mining methods for English reviews are hardly be used for Chinese customer reviews. Now this problem has drawn more attention. J. Liu and others consider sentiment analysis and product features as opinion instance extraction (EIO), and use a supervised method with construct a domain knowledge database and linguistic knowledge database [12]. B. Shi, et al. mining product features on Chinese, but they need to manually create a conceptual model based on product attributes [15]. In Ref. 16 a model is proposed which integrated natural language processing technology with support vector machine. In order to test the performance of the method, they selected 3701 reviews from the 22157 restaurant reviews collection for manual annotation. As the test results, the average precision is 95.6%, the average recall is 81.9% on, and the average F1score is 87.3%.

Above researches are all supervised methods. Ref.6 put forward an unsupervised method and focused on Chinese language character and style to solve above technical difficulties base on M. Hu's mining Algorithm. That paper proposed a method to mining products features from Chinese customer online reviews. The average recall and average precision in their experiments are 77.8% and 63.6% respectively [6]. The work used the words frequent as the main basis for mining. However, the words which appeared in a high frequency may happen to be nothing with the product. This would cause lower precision possibly. Accordingly, we concerns the development of extraction method to adress this problem.

In M. Hu's English reviews mining research, they utilize the observation "opinions tend to appear closely together with features", and find adjacent opinion words that modify the features [2]. Then for each sentence in the review database, if it contains no frequent feature but one or more opinion words, find the nearest noun/noun phrase of the opinion word. The noun/noun phrase is the infrequent feature which needs to be added in the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

feature set. The experiment results have been shown as in table 2.

Table2: Results Of Frequents And Infrequent Features Mining In Hu And Liu's Experiment

Product name	frequent features mining		Infrequent feature identification	
	recall	precision	recall	precision
digital camera 1	0.658	0.825	0.822	0.747
digital camera 2	0.594	0.781	0.792	0.710
mobile phone	0.716	0.828	0.761	0.718
Mp3 player	0.652	0.754	0.818	0.692
DVD player	0.754	0.765	0.797	0.743
average	0.670	0.790	0.800	0.720

In previous research, we deal with the infrequent features in Chinese reviews mostly same as the English way [17]. The results of the experiment (table 3 and table 4) have shown that all the recall values have risen and all the precision values have dropped. The reason is that the supplementary infrequent features might not be real features, and resulted in a lower precision. From the F-score of experiment it could be seen that the complementary infrequent features make overall performance decline.

Therefore, one primary purpose of this work is to seek appropriate algorithm to consider product infrequent features, then to further proke into a efficient unsupervised way. In the next section the proposed algorithm of mining frequent and infrequent features is described. And section 3 shows experimental study for verifying the proposed method.

Table3: Results1 Of Experiment With Infrequent Features Added Directly

Product name	frequent features mining		Infrequent feature identification	
	recall	precision	recall	precision
mobile phone	0.689	0.633	0.750	0.498
digital camera 1	0.805	0.611	0.903	0.437
digital camera 2	0.658	0.641	0.794	0.511
Mp3 player	0.824	0.667	0.856	0.523
Book	0.917	0.629	0.953	0.563
average	0.778	0.636	0.800	0.520

Table4: Results2 Of Experiment With Infrequent Features Added Directly

Product name	frequent features mining	Infrequent feature identification
	F-score	F-score
mobile phone	0.66	0.664
digital camera 1	0.695	0.486
digital camera 2	0.649	0.644
Mp3 player	0.737	0.611
Book	0.746	0.591
average	0.7	0.599

2. METHODOLOGY

In this section, we described our method to mine product infrequent feature from online customer reviews.

The performance of current approach based on association rules mining method shows that the recall is relatively higher, and precision is relatively lower. The infrequent features are extracted as candidate product features directly, which will undoubtedly increase or keep the recall value. Howevery, in fact the additional infrequent features may not be real, and then the precision might decrease. Such the difference between recall and precision will be increased, and the overall performance might be declined. To obtained better performance, the mining method should make the precision value not become lower or stable.

The measures which are used to mine product features from customer reviews follow closely with the concern of customers. As is known, there always exist some features on the various e-commerce Websites or forums about products, but these features may not be paid attention to by common customers(Here we call these feature as static features). And there is little user's sentiment classification on them. It occurs to us to add infrequent features which are collected by using opinion words and further filtered by those static features. This paper proposes the following method and takes an experimental research. The method which focus on Chinese review mining process infrequent words based on associate rules, then modified and expanded. Main ideas are as shown in fig1.

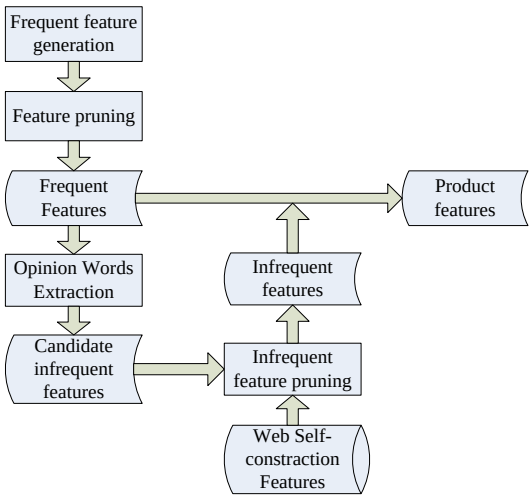

Fig. 1 Main Idea Of The Method For Infrequent Features Extraction

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1 Method Of Mining Frequent And Infrequent Features

We extract frequent features and infrequent features using the following procedure:

Step 1: Parse the reviews corpora (customer reviews downloaded from the Internet). We use the parse tool of Chinese Lexical Analysis System.

Step 2: Identify noun and noun phrases in the input text after parsing by a POS tagged tool.

We also used the POS tagging tool of ICTCLAS (Chinese Lexical Analysis System, Institute of Computing Technology). This POS tagging method has a two levels pattern. The differences are that level one marks only the noun and verb; whereas level two marks more complicated situations such as adjectives or verbs with noun words function, proper nouns, and morphemes. Here in order to increase precision, we used the level two. After POS tagging, basic noun phrases are extracted in accordance with basic noun phrase pattern (as shown in table5) in Chinese customers' reviews.

Step 3: Create a transaction file by reviews after POS tagging.

This step firstly needs to create a transaction database. In this work, a database is saved as a text file. All noun words and basic noun phrases in each sentence are extracted as a transaction in order to find frequent itemsets data in following step.

Step 4: Find frequent itemsets as a candidate features collection I0 based on Apriori algorithms.

Table 5: Patterns Of Tags For Extracting Chinese Basic Noun Phrases

	First Word	Second Word	Third Word
1	Noun	Noun	Not Noun
2	Noun	Noun	Noun (include Location term)
3	Noun	Auxiliary "de"	Noun (include Location term)

This step applies the Apriori algorithms on the transaction file generated in the previous step to find frequent itemsets as the candidate features of products. The minimum support is 1% (refer to the English algorithm). Since three or more frequent items are obviously not the product features, it is same as in English. After exprimental research we have done before, more than three frequent itemsets will not to be considered.

Step 5: Prune I0 into candidate product features sets I1 with compactness rules.

With reference to the definition of compactness rules in English, the compactness rules in Chinese can be defined as:

Let f be a frequent feature phrase and f contains n noun words (or noun phrase). Assuming that a sentence s contains f and the sequence of the noun words (or noun phrase) in f that appears in s is: w1, w2… wn. If the word distance in s between any two adjacent noun words (or noun phrase) and (wi and wi+1) in the above sequence is no greater than 3, then we define f is compact in s.

In this step, traverse every noun phrase, 2 and 3 frequent itemsets f, if f occurs in m sentences in the review database, and it is compact in at least 2 of the m sentences, then we add f to the candidate feature sets I1.

Step 6: Prune candidate products feature sets I1 with redundancy rules into features sets I2。With reference to the definition of P-support of English, the p-support of Chinese can be defined as:

P-support of feature ftr is the number of sentences that ftr appears in as a noun or noun phrase, and these sentences must contain no feature phrase that is a superset of ftr.

In this step, we set minimum p-support to be 3. In other words, if the p-support of a candidate feature is lower than 3, this candidate feature will then be deleted. Filtering out the entire candidate features which p-support is lower than 3, I2 then becomes candidate feature sets I3.

Step 7: Establish common noun words collection in which all words are frequent in Chinese text but not features of products. I3 will be filtered with this collection to be I4.

There are some common oral noun words, such as "difang (place)", "dongxi" (thing)

There are some common noun words to call people, such as "pengyou" (friends), "xiansheng" (sir)

Step 8: Remove the candidate itemsets from I4 to form candidate features collection I5 when the itemsets just contain single Chinese characters. This includes the n frequent itemsets (n<3) and noun phrase which contains single Chinese character noun words.

Step 9: Generate the opinion words. We use frequent features to find adjective words that modify the features. The adjective word which is nearest to candidate feature in I5 will be considered as a subjective opinion words approximately. Then

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

we establish opinion database O1 by using all these subjective words.

Step 10: Use the known opinion words to find those nearby features that opinion words modify. In step 9 and 10, we utilize Hu's observation "opinions tend to appear closely together with features".

Try to traverse all reviews by each sentence. For each sentence in the review database, if it contains no frequent feature in I5, we will try to find whether any subjective word in this sentence matches the element of the opinion words database O1. If a subjective word of the sentence is in O1, the noun or Chinese basic noun phrase which is closest to the subjective word will added to the infrequent candidates datasets IF1, as a candidate infrequent features.

In this step, when we process each sentence in the reviews, there may be two conditions: If only one noun word in the sentence, the noun word will be compliment to IF1 as infrequent features; if the extracted words are of the pattern in Table 5, this Chinese noun phrase will be an infrequent feature to be added to IF1. At last, collection of candidate features which includes the candidate frequent features and infrequent features will constitue the candidate features datasets IF1.

Step 11: Crawler the features from e-business website, then a self-construct dataset F1 will be formed by the static product features.

Step 12: Compared with F1 and IF1, if the element in IF1 would be also in the F1, than we add the element to the I5, otherwise we do nothing. At last, we have the feature datasets I6, which are the colletion of identified frequent and infrequent features in Chinese customer reviews.

2.2 Experiment

A. Datasets

There is no ready-to-use Chinese online customer reviews data on the web. We selected and labeled the reviews download from the web to make the corpus for this study. The corpus of mining experiment are products reviews of products including mobile phones, two digital cameras, a Mp3 player a book. The reviews of digital products are downloaded from IT168 website (http://www.it168.com), and the reviews of book are downloaded from Amazon of China website (http://www.amazon.cn). There are 5 categories of products.

We select 100 reviews for every kind of products and manual tag the products features referred. According to the most minimal coverage principle, we established a minimal products features collection and make this collection can cover all 100 reviews.

We conduct a system which can extract the editor's features from the websites which has the parameters of the products. The system automatically generates a filter collection containing static product parameters which we consider as the static features of products. For digital products, we crawler the website it168 (http://www.168.com), and for books, we use the introduction of product pages on Amazon (www.amazon.cn).

B. Performance indicators

To obtain the performance of the mining method, this study used indicators in the research of text processing, including recall and precision.

In this work the problem of research is to determine whether the result of mining algorithm is the real product features or not, which can be considered as binary classification. Therefore, the performance can be computed by a 2×2 Contingency Table as shown in Table 6. It is a popular method to measure the performance in text classification technologies. The recall and precision are calculated as Eq. 1 and Eq. 2

$$PRECISION = \frac{A}{A+B} \quad (1)$$

$$RECALL = \frac{A}{A+C} \quad (2)$$

Table 6: A Contingency Table For Experiment Performance

	Number of product features	*Number of not product features*
Product features mined by this work	A	B
Product features not mined by this work	C	D

In most cases there is an inverse relationship between precision and recall. It seems often that one is increased at the cost of reducing the other. In order to find whether the performance is increased or decreased or not, it is a better way to combine both two indicators into a single measure. Here we adopt F-sore to measure the harmonic value of both precision and recall as Eq. 3.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$F-Score=\frac{2\times precision\times recall}{precision+recall} \quad (3)$$

C. Experimental results and Discussion

This study used JAVA to build a prototype system for experiments. The extracted opinion words about Digital Camera has been shown as table 7. The results of features extraction and the performance indicators are shown in table 8 and table 9. In comparison to the approach which is only mining frequent features, the proposed method performs better on recall, precision and F-score. From the results we can see some features of products are not change, because the unsupervised extraction method using the Apriori algorithm which we presented in precious study are very effective, though we only deal with the frequent features. Of course, it is obvious that in average the method presented in this paper acquires more plausible performance.

Table : 7Example Dataset Of Opinion Words

Product name	*Opinion words*
Digital camera	teshu(special), shiji(real), da(big), liang(bright), jiandan(simple), dui(right), jingmei(exquisite), xin(new), lao(old), hao(good), duo(many), chang(long), quan(full), leisi(same), anquan(safe), zhenggui(standard), qingbao(light and thin), jingzhi(delicate), buhao(not good)

Table8: Results1 Of Experiment With Infrequent Features Added By Proposed Method

Product name	*frequent features mining*		*Infrequent feature identification*	
	recall	*precision*	*recall*	*precision*
mobile phone	0.689	0.633	0.689	0.633
digital camera 1	0.805	0.611	0.805	0.611
digital camera 2	0.658	0.641	0.722	0.667
Mp3 player	0.824	0.667	0.848	0.700
Book	0.917	0.629	0.917	0.629
average	0.778	0.636	0.796	0.648

Table 9: Results2 Of Experiment With Infrequent Features Added By Proposed Method

Product name	*Frequent features mining*	*Infrequent feature identification*
	F-score	*F-score*
Mobile phone	0.66	0.66
Digital camera 1	0.695	0.695
Digital camera 2	0.649	0.693
Mp3 player	0.737	0.767
Book	0.746	0.746
Average	0.7	0.712

3. CONCLUSION

Within the method of mining product features from Chinese reviews based on associate rule, it is considered to only extract frequent product features. Therefore how to identify infrequent features is a necessary and difficult issue. This work attempts to resolve this dilemma, and we propose an unsupervised method to identify useful infrequent features in Chinese customer reviews. Final experiments verified the effectiveness of the method.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation of China (771001023) and the Fundamental Research Funds for the Central Universities (DL11BB25).

REFRENCES:

[1] Deloitte & Touche: *Industry Outlook*, http://www.deloitte.com. Retrieved May 28, 2008

[2] M. Hu, B. Liu, *Mining Opinion Features in Customer Reviews*, In AAAI Conference on Artificial Intelligence (Page: 755 Year of Publication: 2004 ISBN: 0-262-51183-5).

[3] A. M. Popescu, O. Etzioni, *Product Features and Opinions from Reviews*, In Proceedings of HLT-EMNLP Vancouver, ACL(Page: 339 Year of Publication: 2005)

[4] P.D. Turney, M.L. Littman, *Measuring Praise and Criticism: Inference of Semantic Orientation from Association*, ACM Transactions on Information Systems, vol.21, n.4, pp.315–346, 2003.

[5] E. Riloff, J. Wiebe, T. Wilson, *Learning Subjective Nouns using Extraction Pattern Bootstrapping*, Proceedings of the Seventh Conference on Computational Natural Language Learning(Page: 25 Year of Publication: 2003)

[6] S. Li, Q. Ye, Y.J. Li, L. Rob, *Mining Features of Products from Chinese Customer Online Reviews*, Journal of Management Sciences in China, vol.12, n,2, pp.142–152, 2009.

[7] B. Liu , Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Ch. 11: Opinion Mining, Second Edition(Springer, 2011)

[8] T.L. Wong, W. Lam, *Learning to extract and summarize hot item features from multiple auction Web sites*, Knowledge and Information Systems, vol.14, n,2, pp. 143–160, 2008

[9] M. Hu, B. Liu, *Mining and Summarizing Customer Reviews*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(Page:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

168 Year of Publication: 2004 ISBN: 1-58113-888-1)

[10] L. Zhang, B. Liu, *Identifying Noun Product Features that Imply Opinions*, ACL-2011 (Page: 19 Year of Publication: 2011 ISBN: 978-1-937284-43-5)

[11] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, T. Fukushima, *Collecting Evaluative Expressions for Opinion Extraction*, Proceedings of the 1st International Joint Conference on Natural Language Processing (Page:584 Year of Publication: 2004 ISBN: 3-540-24475-1)

[12] J. Liu, G. Wu, J. Yao, *Opinion Searching in Multi-Product Reviews*, Proceedings of the Sixth IEEE International Conference on Computer and Information Technology (Page:25 Year of Publication: 2006 ISBN: 0-7695-2687-X)

[13] G. Qiu, B. Liu, J.J. Bu and C. Chen, *Expanding Domain Sentiment Lexicon through Double Propagation*, Proceedings of the 21st international jont conference on Artifical intelligence (Page:1199 Year of Publication: 2009 ISBN: 978-1-57735-426-0)

[14] Y.Y. Zhai, Y.X. Chen; X.G. Hu, P.P. Li, X.D. Wu, *Extracting Opinion Features in Sentiment Patterns*, 2010 International Conference on Information Networking and Automation (ICINA) (Page: V1-115 Year of Publication: 2010 ISBN: 978-1-4244-8104-0)

[15] B. Shi, K. Chang, *Mining Chinese Reviews*, Proceedings of the Sixth IEEE International Conference on Data Mining – Workshops. (Page: 585 Year of Publication: 2007 ISBN: 0-7695-2702-7)

[16] C.M. Yu, Mining Product Features from Free-Text Customer Reviews: An SVM-Based Approach, 1st International Conference on Information Science and Engineering (ICISE) (Page: 900 Year of Publication: 2009 ISBN 978-0-7695-3887-7)

[17] S. Li, G. Lu, *Research on infrequent features extraction from Chinese reviews*, Advanced Materials Research, vol. 268-270, pp 1647-1652, 2011.