# Association Rules Based Short Text Feature Extension

**HUANG Wei, Li Shan-fei, Tan Yue-jin, Gao Bing**

School of Information System and Management, National University of Defense Technology, Changsha, China

**Summary**
Focused on the effect on classification of short text sparse features, propose a method extending the short text features. First, according to the theory of words co- occurrence model, the association rules between feature items of corpus are mined by FP-growth algorithm. Then, we search the rules in the set of association rules, which have the relationship with short text feature items, calculate the mutual information between the antecedent and subsequent of association rules, and estimate the degree of association between two features. Based on these work, we choose short text extension feature words and construct the collection of the short text features. Experiments show that the efficiency of short text classification is improved after extending the short text features.

***Key words:***
*Association Rules; Short Text; Text Feature; Extension*

## 1. Introduction

Short text is widely present in our work and life, such as mobile phone short message, web reviews, product descriptions, a summary of a long text, etc, these short text are usually less words, and there is sparse characteristic of each text[1,2], so the measure of similarity between the short text information is not rich, and has brought a certain effects on the short text classification. Therefore, it is needed to extend the feature while using the traditional text classification methods to classify the short text, by adding features to enhance the text feature information.

In literature [3] the author proposed feature co-occurrence set based on extend test texts feature, and to use the way of semantic analysis to reduce the redundancy of feature extended information, but its need for external semantic analysis tools, means and methods is more complex. In literature [4] the author aiming at the problem for text keywords selected has introduced the concept of word clustering, by calculating the association between the two words to determine the clustering, whereas this method is used to extend, its amount of calculation is undoubtedly enormous. However, the existing literature shows that, according to the association between words, the relationship between the feature items of the short text extension is a viable approach. In order to improve the efficiency of a short text feature extensions and accuracy, this paper propose a kind of association rules based on the short text feature extension method, using multi-step strategies for dealing with the association analysis of feature items: first, through the association rule mining to establish the association between the feature items, and then calculate the mutual information values of associated items .At last, following step 2 give the final judgments under the extended feature sets.

## 2. Association Analysis of The Text Features

On short text feature extension, first, it should make define the scope of its extension, namely, the corpus that supported feature extend, and then mining the association rules of corpus between the words. According to the principle of word co-occurrence model[5,6], if in a large-scale corpus, the two words often appear together in the same window unit in the text set (such as a word, a natural segment or a text, etc.), then we judge that this two words are associated, and a co-occurrence probability is higher the more closely associated. The so-called association refers to the values of two or more variables, or there is some regularity between the activities of these association laws, which can be defined as association rules, association rule is used to describe the knowledge model of the law of co-occurrence in a transaction.

If we set the text features of the word as the item, the text for the transaction, we can use the association rule mining method to found the short text feature items in the text corpus and other items of the item co-occurrence relations. Assume item $A$ and $B$ co-occurrence to meet the minimum support and minimum confidence threshold, which can be expressed as association rules $A \rightarrow B$, called $A$ is antecedent of association rule, $B$ is subsequent of association rule.

The existing association rule mining algorithms are mostly based on Apriori algorithm who can generate a large number of candidate item sets when generating association rules. In order to avoid generating candidate item sets, Han has proposed FP-tree based generate frequent item sets, namely FP-growth algorithm, in which, firstly change the problem of mining frequent items into mining FP tree problem, and then use the FP tree item sets for mining frequent item sets[7,8].

Perform the FP-Growth algorithm, in accordance with the minimum support threshold for first pass filtering, and then follow the minimum confidence threshold generation

Manuscript received October 5, 2009
Manuscript revised October 20, 2009

rules to get the frequently co-occurrence feature set[3], and according to the co-occurrence relationship of it to establish the association between feature items in short text and items of the corpus.

## 3. Filtering of Association Rules

### 3.1 Effectiveness of Association Rules

By association rule mining, it has reduced the short text feature extended range, but in the generated association rules there are still some invalid rules, these rules affect association analysis of the text features between each other, which should be initial filtered after association rule mining.

First, the rules should be filtered which antecedent and subsequent of association rules are both in the short text feature set, those rules does not help for the short extension of the text. Assume that a short text with features to be extended is $S$, the association rule is $t_i \to t_j$, if $t_i \in S \wedge t_j \in S$, the rule $t_i \to t_j$ is invalid; if $t_i \notin S \vee t_j \notin S$, rule $t_i \to t_j$ is valid:

$$R(t_i \to t_j) = \begin{cases} 1 & t_i \notin S \vee t_j \notin S \\ 0 & t_i \in S \wedge t_j \in S \end{cases} \quad (1)$$

To obtain the more efficient association rule, association rule lift is calculated after the filtered by equation (1).
The lift of association rule $t_i \to t_j$ is represented as:

$$L(t_i \to t_j) = P(t_j \mid t_i) / P(t_j) \quad (2)$$

If $L(t_i \to t_j) > 1$, say, $t_i$ and $t_j$ are positive association, which means one's occurring implies occurring up of another; if $L(t_i \to t_j) = 1$, say, $t_i$ and $t_j$ are mutually independent; if $L(t_i \to t_j) < 1$, which means $t_i$ and $t_j$ are negative association. Therefore the association rule with the lift larger than 1 should be reserved:

$$R(t_i \to t_j) = \begin{cases} 1 & L(t_i \to t_j) > 1 \\ 0 & L(t_i \to t_j) \le 1 \end{cases} \quad (3)$$

### 3.2 Degree of Association of Feature Items

Association rules is a good example of the association between the feature items, but not all valid association rules can be applied to a short text feature extension, there are also some redundant information, that is, despite the association between feature items, but its degree of association was not high, in which case the extended features of the short text of the classification will become a sort of noise. Therefore, presented paper calculates the degree of association between the feature items using mutual information. According to mutual information of the rule antecedent and subsequent, degree of association between the feature items is calculated, which in order to determine the effectiveness, this approach in dealing with the Chinese text has a good performance [11].
For the antecedent and subsequent part of associate rule $t_i$ and $t_j$, the co-occurrence probability of $t_i$ and $t_j$ is $P(t_i, t_j)$, the probability of occurring up of feature item $t_i$ is $P(t_i)$, the probability of occurring up of feature item $t_j$ is $P(t_j)$, the counts number of occurring up of these feature items in the text are $n(t_i)$, $n(t_j)$, $n(t_i, t_j)$, $n$ is the whole number, so the equation is:

$$P(t_i, t_j) = \frac{n(t_i, t_j)}{n},\ P(t_i) = \frac{n(t_i)}{n},\ P(t_j) = \frac{n(t_j)}{n} \quad (4)$$

The mutually information between feature items $t_i$ and $t_j$ is define as:

$$I(t_i, t_j) = \log_2 \frac{P(t_i, t_j)}{P(t_i) P(t_j)} \quad (5)$$

Mutual information reflects the degree of correlation between the feature items $t_i$ and $t_j$. If $I(t_i, t_j) \ge 0$, that is, $P(t_i, t_j) \ge P(t_i) P(t_j)$, then $t_i$ and $t_j$ is a positive correlation, with $I(t_i, t_j)$ increase, the degree of correlation increased. If $I(t_i, t_j)$ is greater than a given threshold $\varepsilon$, then $t_i$ and $t_j$ closely related. Assume $t_i$ is the feature items of short text, then $t_j$ could be a extension feature of the short text, otherwise it would not be extended.

## 4. Feature Extension of The Short Text

In the process of feature extension in short text, firstly we perform FP growth algorithm for mining association rules in the corpus, gain the frequent co-occurrence feature set, set up the association relationship between the items from the original feature set of the test text and the other items, and then calculate the mutual information value between the antecedent and the subsequent of association rules, set the mutual information threshold as $\varepsilon$ according to the sample size, select the items above the threshold as extended feature of the short text, and finally combine the original feature set with the extension feature set, that is, to obtain the feature set of the short text.
The algorithm of short text feature extension is described as follows:
Input: text set for training, original feature set for testing short text.
Output: feature extension set for testing short text.
Step1: Read in the text set for training, set the minimum support and minimum confidence value, perform the FP growth algorithm, and then obtain the associated rules set Rules{};

Step2: Read in the original feature set Feature{} for testing short text, mark the attribution of all items as Originality;
Step3: Scan the items in Feature{}, for the item $t_i$ which is mark as Originality, perform Step 4;
Step4: Scan the items in Rules{},if there is no association rule including the item $t_i$ ,set its attribution as Extension in Feature{} and save it in the Ex-Feature{};if not, for all the association rules including the item $t_i$ , perform the formula (1)(2)(3), and then execute Step 5 when $R(t_i \to t_j)=1$;
Step5: set the threshold $\varepsilon$ , for the rule $R(t_i \to t_j)=1$, perform the formula (4)(5), then save the $t_i$ and the $t_j$ in the set Ex-Feature{} when $I(t_i,t_j) \geq \varepsilon$ , and mark the attribution of as Extension in Feature{}
Step6: The feature extension end until all the items in Feature{} are marked as Extension. Output the feature extension set of the short text Ex-Feature{}.

## 5. Analysis of experimental result

In this paper, we take 183 capability texts in the version of the U.S. Joint Capability Area as the experimental object. Each one of them should be a short text. After the adoption of the text pre-processing algorithm presented in this paper extend the text features; however, the amount of feature items has an increase of 84.07%.
In order to examine the effect of the text classification after text feature extended, this paper use KNN algorithm [12] to classify the texts before and after the feature extended, classification results of the evaluation index are the precision and recall. This paper classify the Joint Capability Area into six categories, including the joint combat capability of a total of 65 capability as a test text, the remaining 5 categories of a total of 118 capability of training text. Experiment also compared the impact of calculate mutual information to short text classification, results as shown in Table 1.

Table 1 Classification performance comparison before and after Short Text Feature expansion

| | ***pre-feature extended*** | ***pre-calculate mutual information*** | ***calculate mutual information*** |
|---|---|---|---|
| precision | 67.69 % | 70.77% | 76.92% |
| recall | 72.13 % | 70.39% | 74.63% |

Experimental results show that the classification performance has been significantly increased after the short text feature extended, but without accounting the mutual information between feature items, the recall rate of text categorization reduced 1.74%, this is because of the largely redundant rules who has brought noise impact on classification.

## 6. Conclusion

Short text has the characteristics of short length and feature sparse, while using the traditional text classification methods to classify the short text; it is needed to enhance its feature information, that is, the text feature extended. In response to this question this paper proposed an association rules based short text feature extension method, this method can find the association rules between feature items in corpus and text feature items according to the principle of word co-occurrence feature, through filtering association rules to improve feature extended accuracy. Experimental result shows that this method can effectively extend the amount of a short text feature, after feature extended short-text classification performance is can greatly improve. However, the current classification accuracy and recall rate is less than perfect results. The next step will use this method in a more large-scale data environment for further validation and improvement, at same time, improve short-text classification method and further improve the performance of short-text classification.

## References

[1] TENG Shao-hua. Chinese word segmentation and short text classification techniques based on CRFs. Beijing: Tsinghua University, 2009
[2] YAN Rui, CAO Xian-bin, LI Kai. Dynamic assembly classification algorithm for short text. Chinese Journal of Electronics, vol. 37,pp.1019-1024
[3] WANG Xi-wei, FAN Xing-hua, ZHAO Jun. Method for chinese short text classification based on feature extension. Journal of Computer Applications, vol.29,pp.843-845
[4] CHEN Jiong, ZHANG Yong-kui. Novel Chinese Text Subject extraction method based on word clustering. Journal Computer Application,vol.25,pp.754-756
[5] RAK R, STACH W, ZAIANE O R, et al. Considering re-occurring features in associative classifiers. Proceedings of PAKDD, LNCS 3518. Berlin : Springer, pp.240-248, 2005 : .
[6] BAYER T, RENZ I, STEIN M, et al. Domain and language independent feature extraction for statistical text categorization. Proceedings of the

Workshop on Language Engineering for Document Analysis and Recognition. Sussex, UK, pp.21-32, 1996
[7] YI Tong, XU Bao-wen, WU Fang-jun.A FP-Tree based incremental updating algorithm for mining association rules. Chinese Journal of Computers,vol.27,pp.703-710
[8] Han J. et a1. Mining frequent patterns without candidate generation. In : Proceedings of the 2000 ACM SIGM OD Conference On Management of Data. Dallas. TX, pp.1-12, 2000
[9] MA Guang-zhi, ZHANG Sheng-ting. Classifying web document based on association rules. Computer Engineering and Design, vol.26,pp.2515-2518
[10] Sun Maosong, Shen Dayang, Benjamin K Tsou. Chinese word segmentation without using lexicon and handcrafted training data. Proceedings of the 36th Annual Meeting on Association for Computational Linguistics ; Montreal : Association for Computational Linguistics, pp.1265-1271, 1998
[11] DAI Liu-ling, HUANG He-yan, CHEN Zhao-xiong. A Comparative study on feature selection in chinese text categorization. Journal of Chinese Information Processing,vol.18,pp. 26-32
[12] Dasarathy B V. Nearest Neighbor(NN) norms: NN pattern classification techniques.LasAlamitos, California: IEEE Computer Society Press,1991

**Huang Wei** received the M.Sc.degree in higher education research from the National University of Defense Technology (NUDT), China. He is currently a doctor in management science and engineering. His main research interest is system of systems requirements modeling technology, systems management and comprehensive integration technology, data mining.

**Li Shanfei** received the B.Sc.degree in management engineering from the National University of Defense Technology (NUDT), China. He is currently a master in management science and engineering. His main research interest is system of systems demand modeling technology, system management and comprehensive integration technology.

**Gao Bing** received the B.Sc.degree in management engineering from the National University of Defense Technology (NUDT), China. He is currently a master in the technology economics and management. His research interest is technology development programs evaluation research of system-of-systems.

**Tan Yuejin** received the M.Sc.degree in system engineering and mathematics from the National University of Defense Technology (NUDT), China. He is currently a professor, doctoral tutor, and Dean of College of Information Systems and Management in NUDT. His main research interest is system management and comprehensive integration technology, equipment acquisition and project management, complex system theory, and system engineering.