A Novel RS Attribute Reduction Technique for Chinese Classification in CQA

Liwei YUAN, Lei SU*, Peng SHU, Yiyang Li

School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650093, China
* Corresponding author. Email addresses: s28341@hotmail.com (Lei SU).

***Abstract* — with the rapid development of the question and answer services which are based on community, like Sina iAsk, Baidu Zhidao and Yahoo! Answers, the Community-based Question Answering, CQA, service has become a new knowledge-sharing model, which has characteristic of interactivity, openness etc. Increasing more people use the services which are provided on these sites to meet the information needs. In order to accurately understand the user's query and provide useful information, it is necessary to deal with the questions in the community, and the Question classification in the CQA is the key component in this step. However, the difficulty of the question classification is High-dimensional feature vector, which usually uses the feature selection as the primary method of dimensionality reduction, and in this paper, a combined extract features method is presented to screen the question features for the first time to obtain a feature subset. Then the importance and dependence of the rough set is used as the heuristic information for the feature selection to further screen the useful features. Experiments show that the algorithm is effective, and it not only make the question of feature dimensions reduced to some extent, but also improve the classification accuracy of the question.**

Keywords - Community-based Question Answering; Question classification; rough set; Extract Features; Attribute reduction.

I. INTRODUCTION

At present, the common CQA service includes Baidu Zhidao, Sina iAsk, SoSo, Yahoo!Answers in China community-based QA. As a typical interactive application, it is more and more popular to get the online information with the community-based question and answer service. The user in the community is not only a consumer of online information, but also the provider of the information, because the community-based QA makes the user share their knowledge and experience together [1]. Now, community-based question and answer service has gradually been paid attention by the industry peers and the researchers in related fields, which is becoming a hot topic.

Question Classification is an important part of a cQA system, which is the basis to develop the strategies of taking answers and pinpointing answers, and the accuracy of classification directly affects the performance of question answering system [2]. The usual task of question classification is to classify the question by the result that the user desired, such as the type of question aiming for the data, the types for names and so on, to guide the auto QA system to find answers to the questions. But in cQA service, the purpose of classification task is to assort different questions into different topic field, such as business, tourism, computer and network, which is a kind of automatic text classification. The statistical learning algorithms of the usual text classification mainly include the Naive Bayes (NB) [3], kernel method[4], Snow[5], KNN（Nearest Neighbors NN）, and SVM, etc. It is different from the traditional distance metric method between document and categories that Bigi presents a new distance metric method: Kullback Leibler Distance [6]. The method can measure the distance between the word probability distribution and various types of word probabilities in the classified documents to guide the classification algorithm to put the classified documents into the closest category in its distribution. Besides Tian Weidong et al. [7] adopted a policy of self-learning rule to classify, which uses POS information to build the rule of "question words: category" and "question word +headword: category" for Chinese Question Classification, and it achieved good classification results in experiment. Singh et al [8] used the language model that is based on the translation to carry out the Question Classification, that language to the vector space, and the effect should be better compared to the vector space model. Questions in cQA are usually short and cannot provide sufficient syntactic and semantic features. To tackle the problem of data sparseness, Hotho et al. [9] used the synonym and hypernym included in WordNet to expend the textual characteristics. Cai et al. [10] proposed a two-stage approach for question classification in cQA. The large-scale categories are pruned to a small subset, and then the questions are enriched by leveraging Wikipedia semantic knowledge (hypernym, synonym and associative concepts).

There are a lot of similarities between Question Classification in cQA and traditional text classification, but also there are many differences. If the existing classification method is applied directly to the cQA, it will lead a significant reduction in classification accuracy. On the one hand, different from the normal texts and documents, the questions in cQA are usually short. Therefore, the traditional learning model which is based on the bag-of-word in vector space model extracts a lot of feature value with zero due to the data sparseness. On the other hand, the original feature space may be provided by all the features in questions. The huge number of feature words, but some features words often not only appear rarely, but also the contained category of information does not have good capability of characterization for the features of questions, , which will increase the dimension of the vector space. Classification algorithm and the complexity of implementation will be increased with the increase of the feature space dimension. So the effective screening of features can not only improve the training time of question classification, but also can improve the accuracy of classification.

To solve the above question of high-dimensional problems, this paper first adopts the similarity between terms and terms to obtain the feature space, and after, the dimensionality reduction of questions feature space is performed by using the attribute reduction based on the rough set, which not only solves the problem of the disaster of the space dimension, but also reduces the complexity of the classification algorithm. The experimental results have demonstrated that the method is effective.

The rest of this paper is organized as follows. Section 2 introduces the method of feature extraction. Section 3 introduces the attribute reduction based on Rough Set. Section 4 reports and analyzes the experimental study on the Chinese question classification in cQA. Finally, section 5 summarizes this paper and introduces the future work.

II. METHOD OF FEATURE EXTRACTION

In process of feature extraction, the word segmentation plays a crucial role. ICTCLAS platform (http://ictclas.nlpir.org/) is used to do word segmentation for Chinese questions and stop words are removed in this paper. When constructing the marked category for the artificial collated fields, you will find a lot of damaged entries that induced by the word segmentation tool, which is mainly due to the lack of appropriate rules for the user's dictionary. In word segmentation, there are many feature words that has obvious differential ability to such categories, , because the problem that caused the destruction of parts of the speech, for this problem, Bin Yang proposed to build a proprietary user dictionary [13] to solve the problem of part speech in the process of word segmentation.

A. Domain Proper Nouns Table

In Community-based Question Answering system, there are some domain proper nouns which can contribute to the identification for the category. For example, Dungeon-Fighter is a popular game and becomes a hot topic in the game community of Baidu Zhidao, the domain term "地下城" cannot be spitted into "地下 城". Therefore, the domain proper nouns are collected and adopted as a dictionary for the word segmentation tools.

B. Feature Table upon Category Tree

The categories are organized as multilayered structure in cQA. In fine classification (it is the classification for the layer 2 categories).Because of the use of the feature extraction algorithm, its features will be sorted. The features with the coarse categories are extracted, and the data sparseness is very obvious because of the huge amount of questions. Therefore, the fine features are extracted according to the bottom level. We set a predefined dimension D, and extract the high-frequency terms from the labeled questions. First, the number of terms belong to the category i with high frequency is Counti, and the number of overall terms is counted as Sum. Then, the proportion of the number of the term with high frequency is set to Pi= Counti/Sum. So, the number of features extracted from the labeled questions upon the bottom categories is Ni= Pi*D.

C. Combined Extract Features Algorithm

Currently, the main selection algorithms for the feature words of the text categorization are: DF, IG, MI, CHI, etc. But suppose these feature extractions are all based on assumptions that the feature items are independent, and the focus of each method is different. DF stressed that the effect of the high frequency on question classification, ignored the word which appears less influence to the classification, neglected the effect of the word on the classification. MI's feature selection trends to affect the classification of rare words. Literature [14] by introducing the disturbance factor achieved good results in dimensionality reduction and classification accuracy. But it is not satisfied with the low-dimensional data processing. CHI focuses on the correlation between the entry and the category. The more relevant is, the greater the contribution to the categories. These assumptions will cause the loss of part of classified information, and affect the result of the following Classification question. Aiming at the advantages and disadvantages for each feature extraction algorithm, Citation [15] put forward a united feature extraction method—CEFA, and made corresponding improvement. To combine DF with IG, MI, and CHI to make feature selection, in order to remove redundant features and retain the important feature of the low-frequency word.

D. The Questions Feature Extracting Value

This paper selects the bag of words as the classification features, getting through the proposed

Combined Extract Features Algorithm to obtain classification feature by setting a corresponding threshold, and gets a different number of dimensional vector space model (VSM), due to the questions in cQA are usually short. Therefore, the traditional learning model based on the bag of words in vector space model extracts a lot of feature value with zero due to the data sparseness. Since the word in question appears, although not appears in the feature space, it may also have a strong correlation with the features dimension of the feature space. The question of the words appears in the feature space on a particular dimension, and other features dimension also exist a certain correlation. Therefore, to solve the data sparseness problem of the question classification, the method for calculating the semantic similarity of vocabulary is adopted , for the words that appear in each question, the similarity computation is carried out for the feature dimension words in the feature vector space respectively, and the similarity value-weighted is used for the weighting of features dimension , and the relevance of the words could be reflected in the feature vectors of each questions, to resolve the data sparseness problem of the question .

In order to get the feature vector of each corresponding question, the feature vector space Wk obtained through the statistics is (W1,W2,…,Wk,…,Wn) , and each question Ti can be regarded to be constituted by the independent features entries' group (Wi1,Wi2,…,Wik,…,Win). The features word Wik of each feature in turn calculate the similarity with feature space attribute word Wk through the vocabulary semantic similarity, and the similarity is calculated by using Liu Qun "Based on HowNet semantic similarity is calculated" [16], the formula is as follows (1)

$$Sim\,(W_1, W_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_{\,j}\,(W_1, W_2) \qquad (1)$$

Using Sim1 (W1,W2) as the first sememes similarity, Sim2(W1,W2) as other sememes similarity, Sim3(W1,W2) as the relational sememes similarity, Sim4(W1,W2) as the symbol sememes similarity, βi (1≤i≤4) as the adjustable parameters, where β1 =0.5, β2 =0.2, β3 =0.17, β4 =0.13.

Through the word similarity calculation, the question can in turn get the word and other word-dimensional feature space weights (A1k, A2k,…, Aik,…, Ank). The corresponding word in this question and feature space attribute values is calculated by averaging other weights Aik. The formula is as follows (2)

$$M_{ik} = \frac{\sum_{i=1}^{n} A_{ik}}{n} \qquad (2)$$

In which n is the number of dimensions of the feature space, and the average weights is calculated for each question in the feature space. Put this value as the feature value of this feature dimension of the questions, and regard the feature attribute Wk as the horizontal axis of a k-dimensional coordinate system, , and the feature entry group (Wi1,Wi2,…,Wik,…,Win) of the questions Ti as the vertical axis of the coordinate system , (M11,M12,…,Mik,…,Min) for the corresponding coordinate values. Such a feature that extracted from any of the question constitutes an n-dimensional feature vector, m questions (T1,T2,…,Ti,…,Tm) can be configured to extract a feature vector of dimension m*n.

III. ATTRIBUTE REDUCTION BASED ON RS

Rough set theory [17] is proposed by a professor at the University of Pawlak of Poland in 1982. As the new mathematical tools for processing the knowledge of incompleteness, inaccuracy, uncertainty, the rough set theory is to analyze the data obtained directly that does not require any initial or additional information to be handled in advance, mainly to mine the characteristics for the rule that hidden in the knowledge and to express its form with their knowledge. In fact, the core of the entire mining process is the attribute reduction and rule extraction of the knowledge [18]. Through thirty years of development and improvement, rough set has not only established a relatively complete theoretical system, and also developed the technology model design by the use of rough set theory, and application system is becoming mature. With the current rough set theory continues to mature, the rough set has been successfully applied to the field of decision analysis, artificial intelligence, pattern recognition and text-mining and others.

A. Knowledge representation method of RS

Knowledge representation system of knowledge is to find useful knowledge from large amounts of data through a certain method of describing the order or decision rules by studying the basic characteristics of the value of the specified object. It is possible to use the knowledge representation method of information systems to achieve the accurate knowledge of the rules found in the description and expression by the rough set theory [19]. Features (attributes) knowledge representation system by specifying the object and the characteristics (property values) is described, and a typical method of rough set knowledge representation system S can be described as:

$$D(s) = (U, C \cup D, V, F) \qquad (3)$$

Where: U is the non-empty object, called domain; C, D is the set of attributes condition and results of a set of attributes, is a range of property and an information function, which can be defined for each attribute of an information value, which is:

You can use the data in tabular form for easy and accurate expression of the knowledge rules of the knowledge representation system of the rough set theory. In the data table of the knowledge representation system, the raw of the relational table corresponding to the research object, and each row represents a message of the system and the column corresponding to the object's properties, and the information of an object is expressed through the property value of the specified object, so the each property of the information system will correspond to an equivalence relation, and a table of information is defined as a family of equivalence relations, that is knowledge-based systems.

TABLE 1 DECISION INFORMATION TABLE

U	C（condition attribute）				D（decision attributes）
	C1	C2	C3	...	
x1	2	1	4	...	1
x2	5	4	2	...	2
x3	7	3	6	...	3
...	...	...	...	...	...

Training samples in this article can be converted into a form of expression of this decision table. Here, in order to meet the requirements of rough sets on the treatment methods and procedures that can be performed more easily, some of the original property has been quantified and replaced by discrete values; decision attribute also is quantized into discrete values. There are many kinds of discrete method of the rough set theory, which are used by us in the analysis of the data equidistant. [20] The decision table with discretization.

B. *The importance of calibration characteristics*

The above decision table and select Properties directly affects the classification effect and time back. The calibration of attribute importance refers to the property minus the observed classification changes. If it changes, it indicates that the property is important, on the contrary, it shows that the importance property is low. Calibration procedure has the following:

(1)To calculate the domain of decision attribute D with respect to the condition attribute C in the two-dimensional decision table:

$$Pos(D) = \bigcup_{X \in U/D} CX \tag{4}$$

The domain representation of Decision attribute D with respect to the condition attribute C is the training question set that can be accurately classified to the equivalence class of decision attribute D on the basis of the information of the classification information U / C. If the classification is complete, it should be the entire sample set U.

(2)According to the dependent function of the rough set:

$$K = \gamma_c(D) = \frac{card(Pos_c(D))}{card(U)} \tag{5}$$

Card refers to a set of base requirements. K=1, explains the decision attribute D is completely dependent on the condition attributes C; K<1, explains the part dependence on the decision attribute D; K=0, describes that the decision attribute does not depend on the condition attribute C.

By using this formula to our feature selection, you can know the dependence of the decision attribute D on the condition attribute C is the ratio of all the number of questions in U that can be accurately divided to the equivalence class of the decision attribute D according to the information of the category U / C and the number of the entire sample set. (3)In the decision-making table, each attribute ti, can calculate its importance Sig (ti, C, D) of the decision attribute D:

$$Sig(t_i, C, D) = \gamma_c(D) - \gamma_{C-ti}(D) \tag{6}$$

Decision attribute D represents the dependence on the condition of the property C, refers to the remove of the property ti in the condition attributes C, and the decision attribute D dependents on the condition attribute C. The higher the value Sig (ti, C, D) is, the greater the importance of the property ti to the classification.

C. *Attribute Reduction*

In the property on the basis of the importance of calibration, I = (t1, Sig (t1, C, D)) ... (tn, Sig (tn, C, D))) is the vector of attribute importance. Reduction of property can be achieved by.

(1)In accordance with the size of this array (Sig (tn, C, D)) of the sort, delete the property Sig (tn, C, D)) = 0, because it does not work for the classification.

(2)Set a threshold, delete the property that tn, (Sig (tn, C, D)) is less than the threshold value, and its aim is to remove some of the classification with less important attribute, thereby reducing the time to produce rules in the sample training stage, but the accuracy of the generated rules would be affected. So it should try to make the threshold value of the minimum, to ensure the accuracy of the rule, and the threshold value can be set to zero when the number of the condition attribute is not too much.

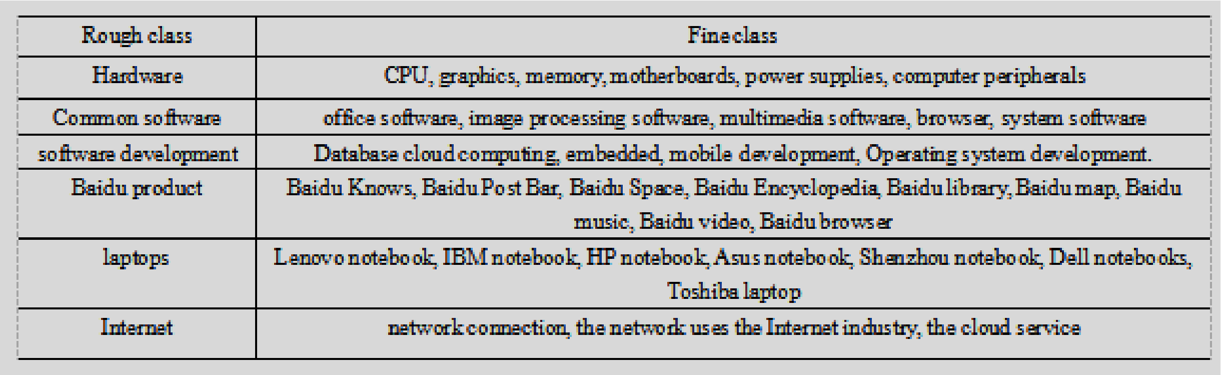
Feature selection based on the feature selection algorithm of the rough set is an attempt to set its rough dependence function to be applied to the feature selection, and this method can reduce the number of the dimensions of the feature set, which has become the method RSFS.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Corpus

Chinese question data collected from Baidu Zhidao are used as the training examples in the experiments. The types in the categories tree can be divided into three layers from top to bottom. The top layer contains 13 categories, and the second layer contains 141 categories. In the second layer, there are 41 categories can be divided into the third layers, and then the third layer contains 289 categories. During the experiment, we select questions from the big category "computer network", a total of 24,561. The questions system is as shown in below Table 2.

TABLE 2 QUESTION CLASSIFICATION SYSTEM



Rough class	Fine class
Hardware	CPU, graphics, memory, motherboards, power supplies, computer peripherals
Common software	office software, image processing software, multimedia software, browser, system software
software development	Database cloud computing, embedded, mobile development, Operating system development.
Baidu product	Baidu Knows, Baidu Post Bar, Baidu Space, Baidu Encyclopedia, Baidu library, Baidu map, Baidu music, Baidu video, Baidu browser
laptops	Lenovo notebook, IBM notebook, HP notebook, Asus notebook, Shenzhou notebook, Dell notebooks, Toshiba laptop
Internet	network connection, the network uses the Internet industry, the cloud service

Experiments used 10-folds cross validation, divided the data set that composed of the 24,561 questions into 10 folds, and each time took a fold as a test set. The test set is 10% of the entire data set, and the remaining 90% of the data composed training set by the nine folds. 10 folds are in turn taken as a test set for testing, running 10-folds cross validation 10 times independently on average.

B. Experimental Process and Analysis of its Results

Experiments consist of two parts. In the first part, after using the field proper nouns, we use SVM classifier to verify whether the effect of the classification is lifting or not .In the second part, after thinning the characteristics of the question on this basis, we use the rough set to conduct the attribute reduction and use SVM, KNN, Naive Bayes three different classifiers training experimental model respectively. We use different classification methods to classify the comparative results of the question. During the experiment, the classification levels are divided into two kinds of classification, the Rough Classification and the Fine Classification. The so-called Rough Classification is about conducting classification experiments for six big categories, and the so-called Fine Classification is about conducting classification experiments for 36 small categories.

B1. Based on Exclusive Vocabularies Classification

In reality, new terms are emerging in large numbers continuously and such terms often represent the specific areas. To compensate for the feature words whose parts of speech are destroyed in the process of segmentation, we introduce the proprietary vocabularies. In the experiments, we use SVM as the classification. Therein, we set the penalty factor as 10, and the parameters are default values.

TABLE 3 INTRODUCES PROPRIETARY VOCABULARIES CLASSIFICATION EXPERIMENT

The Method of Feature Extraction	Dimension	
	100	100
	Rough Classification	Fine Classification
Without Domain Proper Nouns Table	62.1%	47.2%
Domain Proper Nouns Table	64.6%	50.2%
Feature Table upon Category Tree	65.8%	51.2%

From the data, we can see that in the case of not introducing the proprietary vocabularies, the classification condition is 62.1% (Rough Classification) and 47.2% (Fine Classification) respectively. After introducing the field proprietary vocabularies, the classification is 64.6% and 50.2% respectively, which have increased by 2.5% and 3%. Later, through the category of proprietary vocabularies, we improve the classification results, and compared with the above case, it has raised 1.2% and 1.0%. As shown in the above table, we can see that the introduction of the field proper nouns table and feature table upon category tree make the field features and characteristics of the various categories have better retention and the accuracy has been further improved.

B2. Based on Combined Extract Features Algorithm Questions Classification

The traditional methods of feature extraction have their advantages and disadvantages. To compensate for the shortcomings of various feature extraction algorithms, we propose a Combined Extract Features Algorithm to improve the traditional feature extraction algorithm. In the experiment, for the only question categories, we conduct classifying experiments. Classifier is selected as SVM in there, and we set the penalty factor as 10 and the remaining parameters remain unchanged to validate the feasibility of this algorithm.

From the above table, we can see that the effect of classification of the combined extract features algorithm proposed by this paper has a little promotion than the traditional feature extraction algorithms. Compared with the most stable DF algorithm in performance, it relatively increases nearly 1%, so this algorithm is effective.

TABLE 4 QUESTION CLASSIFICATION OF DIFFERENT FEATURE EXTRACTION ALGORITHMS

Feature Extraction Algorithms	Dimension		
	100	200	300
DF	64.8%	65.2%	65.5%
CHI	62.3%	62.8%	63.2%
MI	64.4%	64.9%	65.4%
IG	62.8%	63.0%	63.5%
CEFA	65.8%	66.1%	66.9%

C. Based On Classification Of Attribute Reduction Of Rough Set

In the experiments, after using the field proprietary vocabularies, we solve the problems of destruction of parts of speech in the segmentation process. We use joint feature extraction algorithm proposed by this paper, after setting the appropriate threshold, and we extract the feature subset of 100- dimension, 200-dimension, and 300-dimension respectively. Without destroying the original classification results is the feature of the classifier. On this basis, we refer to the rough set to conduct a second screening for features, and then reduce the feature further to screen for effective feature words of the classification.

Experiments use two evaluation methods: Vector dimensionality reduction rate and classification accuracy rate.

Vector dimensionality reduction rate is the number of invalid features which is reduced divided by the number of before Decision Table Reduction features. Its mathematical formula is as follows:

$$\text{Vector dimensionality reduction rate} = \frac{\text{the number of invalid features which is reduced}}{\text{the number of before Decision Table Reduction features}} \quad (7)$$

The number of features which is reduced is that the number of features after initialization screening minus the number of after Decision Table Reduction features Classification accuracy rate is the ratio which is shown in questions, which are consistent with the manual classification results in all the questions involved in the testing classification. It is used to evaluate the performance of the classification. The mathematical formula is as follows:

$$\text{Classification accuracy} = \frac{\text{The number of correct Categories questions}}{\text{The number of actual classification questions}} \quad (8)$$

TABLE 5 THE NUMBER OF FEATURES OF ROUGH SET FEATURE SELECTION ITEM

Dimension	100 Dimension	200 Dimension	300 Dimension
The Number of Features after Pre-treatment	100	200	300
The Number of Features after Reduction	73	159	234
Vector Dimensionality Reduction Rate	73%	79%	78%

In order to demonstrate the effectiveness of the rough set attribute reduction, we use SVM, KNN, NaiveBayes three different classifiers training experimental model respectively, and verify the classification accuracy respectively. Therein, SVM parameters are the same with the above experimental parameters. In KNN algorithm, the value of K is between 10 and 35. Specific experimental data is as shown in Table 6, Table 7, and Table 8.

TABLE 6 RESULTS OF USING SVM TRAINING MODEL

Categories	Classification Accuracy Rate					
	100Dimension		200Dimension		300Dimension	
	No Reduction		No Reduction		No Reduction	
Rough Classification	65.8%	69.2%	66.1%	70.3%	66.9%	72.4%
Fine Classification	51.2%	54.3%	51.9%	55.1%	52.3%	56.6%

TABLE 7 RESULTS OF USING KNN TRAINING MODEL

Categories	Classification Accuracy Rate					
	100Dimension		200Dimension		300Dimension	
	No Reduction		No Reduction		No Reduction	
Rough Classification	67.1%	70.8%	67.9%	71.4%	68.6%	74.1%
Fine Classification	52.9%	56.1%	53.4%	56.9%	54.1%	57.7%

TABLE 8 RESULTS OF USING NAIVE BAYES TRAINING MODEL

Categories	Classification Accuracy Rate					
	100Dimension		200Dimension		300Dimension	
	No Reduction		No Reduction		No Reduction	
Rough Classification	63.7%	66.8%	64.4%	68.3%	65.6%	69.7%
Fine Classification	50.4%	53.1%	51.9%	54.3%	52.2%	55.8%

The results show that: after the attribute reduction, the classification is increased compared with previous attribute reduction. In SVM algorithm, for example, the use of rough set in 100-dimensional feature vectors represents an unused rough set, which is improved 3.4% in rough classification process. Use the 200-dimensional feature vector in rough classification process improves 4.2%. Use the 300-dimensional feature vector in rough classification process increases by 5.5%. Only by using the attribute reduction of rough set theory can improve the classification results. The traditional rules, after obtaining the attribute reduction, basically contains attribute which is included in the all attributes combinations, that is, these attribute, after reduction, can well replace the text information. Features are more representative. Although the feature extraction has been removed the features which are helpless for the classification, the attribute reduction still further reduces the dimension and improves the classification results.

From Figure 2 and Figure 3, we can see that in the experiments of question classification, the selection of different classifier has different classification result of a particular sample. In this experiment, the classification result of KNN algorithms has significant distinction between SVM and Naive Bayes. Its classification result is better than the other two algorithms.

Compared with the current main methods of question classification, such as SVM, KNN, NaiveBayes, the rough set theory, which is used for classification has the following advantages: 1, it could obtain the classification required minimum attribute features and could reduce dimension of feature vectors under the conditions of not affecting the classification accuracy. 2, it could get classification rules of the most simple explicit expression, and reduce the computational complexity of the classifier algorithm.

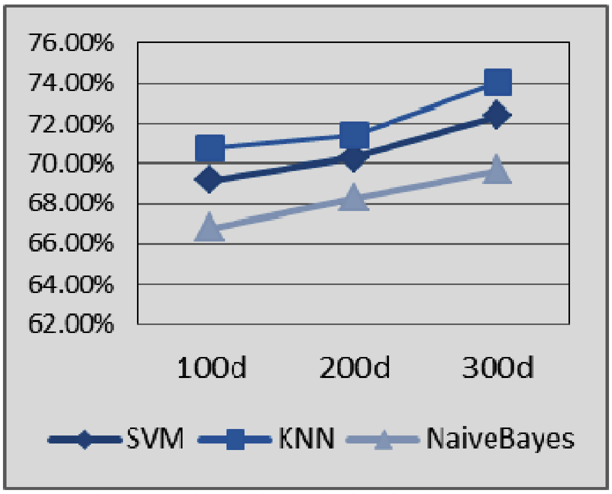


Fig.2 comparison with Rough Classification accuracy

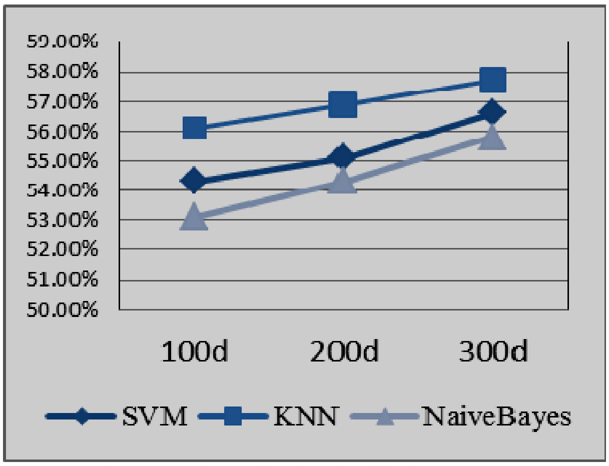


Fig.3 comparison with Fine Classification accuracy

V. CONCLUSIONS AND FUTURE WORK

In this paper, the rough set attribute reduction has been applied to the question classification for its difficulties and challenges existed in the cQA, which not only reduces the spatial dimension effectively, but also reduces the complexity of the classification algorithm, and introduces the field proper noun form, uses the united feature extraction, so that the correlation of the feature and category can be enhanced continuously. Experimental results show that this method can significantly improve the accuracy of the question classification. Researches on cQA have achieved certain results, however, as an emerging research topic, some unresolved issues still need to be studied, and we need to continue to explore and improve.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61365010), Yunnan Nature Science Foundation (2011FZ069).

REFERENCES

[1] Mao Xianling,Li Xiaoming. Survey of question answering system. Journal of Frontiers of Computer Science and Technology,DOI: 10.3778/j.issn.1673-9418.2012.

[2] H. Duan, Y. Cao, C. Y. Lin, and Y. Yu. Searching questions by identifying questions topics and question focus. In ACL, pages 156–164, 2008.

[3] Zhang Yu,Liu Ting, Modified Bayesian model based question classification[J].Journal D Chinese Information, Processing , 19(2):100—105 ,2005

[4] Taira Jun Suzuki,Sasaki Yutaka, and Maeda Eisaku Question classification using HDAG kernel[c].ACL Workshop on Mulitilingual Summarization and Question Answering, Sapporo, 61-68, 2003

[5] Li Xin and Roth Dan Lerning question classifier[C] In Proceedings of the 19th International Conference on Computation

[6] Bigi B.Using Kullback-Leibler Distance for Text Categorization[C] Process of the 25th European Conf on IR Research, 305—311, 2003

[7] Tian WeiDong,Gao YanYing,ZuYongLiang. Classification of problems based on self-learning rule and improved Bayesian combination [J]. Computer application research, 27(8):2869—2871, 2010

[8] SINGH A,VISI_ESWARIAH K CQC: classifying questions in CQA Websites[C].Proceedings of the 20th ACM international conference on Information and knowledge management Glasgow,United Kingdom, 2033—2036, 2011

[9] H0TH0 A,STAAB S,STUMME G. WordNet Improves Text Document Clustering[C]. Proceedings of the Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto Canada, 41-544, 2003

[10] Li Cai, Guangyou Zhou, Kang Liu, and Jun Zhao. Large-Scale Question Classification in cQA by Leveraging Wikipedia Semantic Knowledge. In Proceeding of the 20th ACM Conference on Information and Knowledge Management, 2011.

[11] X. Xue, J. Jeon, and W. B. Croft. Retrieval models for question and answer archives. In SIGIR, 2008.

[12] Y. Cao, H. Duan, C.-Y. Lin, Y. Yu, and H.-W. Hon .Recommending questions using the mdl-based tree cut model. In WWW, 2008.

[13] Su Lei, Yang Bin, Qi Xiangxiang, Xian Yantuan. Refined feature extraction for Chinese question classification in cQA. In ICST. Pages 318-326, 2014.

[14] Liu Jian,Zhang WeiMing. Research and improvement of text feature selection method based on mutual information [J] Computer engineering and Application, 44(10):135.137, 2008

[15] Zhou Cheng,Ge Bin,Tang Jiuyang,Xiao Weidong. Combined feature selection method based on correlation and redundancy degree [J]. Computer science, 39（4）, 2012

[16] Liu Qun, Li Sujian. Word Semantic similarity computation based on HowNet. In: Proceeding of the third Chinese word semantic conference, China Taibei,2002

[17] Pawlak. Z Rough sets[J] International Journal of Computer and information Science, 11:341-357, 1982

[18] Peters J F, Skowron A. Transactions on Rough Sets IV [A].New York:Springger-Verlag,2005.

[19] Wang Guoyin. Rough set theory and knowledge acquisition[M]. Xi'an JiaoTong University.2001.

[20] Liu Yezheng.Jiao Ning.Jiang Yuanchun. A comparative study of continuous attribute discretization algorithm [J]. Computer application research, 24(09):28—30, 2007.

[21] Hong Zhiyong,Liu Hua,Deng WeiBin. New text classification method based on rough set and correlation vector machine [J] Computer simulation. 27(7):1479-184-1, 2010.

[22] Z .Rough sets [J].International Journal of Computer and Information Science, 11:341-357, 1982.