Research on Pretreatment of Questions Based on Large-scale Real Questions Set

Qingsheng Wan, Shaobin Huang, and Mengxi Wei
College of Computer Science and Technology, Harbin Engineering University, Harbin, China
Email: wanqingsheng@yeah.net

Abstract—**Aiming at the set of large-scale real questions collected from the question-answering system based on community, a novel pretreatment method of questions is put forward. The method uses KNN-based active learning algorithm to train the classifier, and uses the neighborhood category system to classify the questions efficiently. On this basis, the classified set of questions is divided into equivalence classes so that the semantically related questions are got together in combination with statistical information, semantic information and subject information based on the LDA model. The final experimental results show the effectiveness of the methods proposed.**

Index Terms—**Question Set, Pretreatment, KNN, Active Learning, Equivalence Class**

I. Introduction

Qquestion-answering system based on community contains a large number of pairs of question and answering. These real pairs of question-answering provide the corpus with great research value for questions analysis, information retrieval and answer extraction of question-answering system. However, many of these real questions are repeated or semantically related, which will undoubtedly affect the analysis, matching and extraction efficiency if they are directly regarded as research corpus of question-answering system. So the questions should be preprocessed, namely the question classification. Question classification is the first crucial processing flow in question-answering system that can effectively reduce the search range and the space of candidate answers, and improve the accuracy of the returning answers of the system. Question classification guides the subsequent module of question-answering system. So the result of question classification directly influences the quality of question-answering system.

The question classification methods mainly consist of the methods based on artificial rules [1] and the machine learning methods based on statistics. The former works based on keywords and regular expressions. However because of the complex polytrope of language, it is difficult to make the rules. The latter has more obvious advantages, and is widely used. It builds the model that is used to categorize and realizes the recognition of unknown question types by machine learning based on the labeled question words in training sets. There are many methods of machine learning, for example, references [2, 3] proposed a question classification method based on support vector machine (SVM), and reference [4] proposed a question classification method based on sparse network of winnow by using the idea of hierarchical classification. However, the above two types of methods both need to label the categories of training set first, to support the further divisions of question categories. For example, The commonly used corpuses for question classification training in Chinese question-answering system are mainly the question set of the question-answering system of Harbin Institute of Technology Information Retrieval Laboratory (HIT-IRL) [5] and the Chinese question classification data set (version 1.0) of Fudan University [6]. The question set of HIT-IRL includes 6264 questions with type information. The question categories cover 6 large classes and 63 small classes, and each class of questions has been revised artificially. The data set of Fudan University contains 17,252 questions. Its question classification system is established by referring to HOWNET, and the data is selected from two aspects of question form and answer type. Each of the questions is described by one <QuestionStyle> property and two <answer of type> properties, and the Chinese problem categories are marked artificially. The two questions corpuses are both the training sets for open fields, and the questions are less complex and small in quantity, so the training demand can be met by hand-annotated. However, in real question sets for practical application domain, the manifestation pattern of questions is more complicated, and different fields may use different classification systems, especially when the scale of data set is larger, hand-annotating is very difficult. For example, in the pension insurance question-answering system studied in this paper, 140,000 related questions in real fields proposed by users are got from Baidu Know and Tencent Soso, and these questions cover many subject categories such as the insured, payment, personal account, relation transferring and continuing, treatment development and adjustment, enterprise annuities, social management services and so on. It will undoubtedly consume a great deal of time and effort to build training set by hand-annotating the question categories.

Consequently, in order to reduce the pressure of artificial labeling, this paper proposed a two-stage question classification method. Firstly, we pre-classify the questions by using the question classifier based on KNN active learning, and a small part of questions are

doi:10.4304/jnw.8.8.1810-1816

extracted from the question set to be marked manually. Then more useful data for classified model is selected to be marked from the sample set that will be marked by using the heuristic method in order to expand labeling sample set. Later on, the iterative learning continues in the new marking and to be marked sample sets so as to train the higher accurate classification model and produce the final classifier. Subsequently, the classified question set is divided into equivalence classes and semantically related questions are organized together in combination with statistical information, semantic information and subject information based on LDA model. By the two steps mentioned above, question set is organized into a suitable form that is used by question and answering research.

The remaining of this paper is organized as follows. We discuss related work in Section 2, namely the question classification methods based on KNN active learning used to categorize the questions initially. In Section 3, we propose the definition and method about equivalence classes division. Experimental results are presented in Section 4, which indicates the accuracy of our methods, and the effect in basic endowment insurance domain also indicates the validity of our method. Finally, this study is concluded in Section 5.

II. Related Work

A. Active Learning

Traditional machine Learning belongs to passive learning, which takes the already tagged sample set given outside as the training sample, and generates classifier by accepting sample information passively. Although it can get a better classifier, the passive learning algorithm not only needs a great number of training sample sets, but also demands each training sample marked artificially which will consume time and effort.

Reference [7] proposed active learning algorithm to solve this problem. The active learning does not randomly select and artificially mark the training samples, but is able to actively choose those samples that contain more information and are more helpful for classifier performance, and these samples are transferred to training set for further training to generate a new classifier.

Compared with those methods that randomly select samples which are needed to be marked and trained, the classifier produced by active learning algorithm can identify some sample subsets containing more information from a large number of unlabeled samples each time. So labeling personnel don’t need to take time to mark those samples which are not helpful or less helpful for further training samples, thus it is able to avoid the waste of resources and effectively reduce the cost of getting the training sample. At the same time, because active learning can control the size of training sample set, the computing scale of classifier construction is also greatly reduced [8].

Active learning is a circulating process in the form. Firstly, mark two data sets, namely the training data set L and the test data set T. The initial training data set L only contains a small amount of samples, and unlabeled samples are all assigned to unlabeled data set U. The process of active learning is as follows: (1) Use the samples of the training data set L to learn a classification model; (2) Use the classification model to predict the samples of the unlabeled data set U, and select the first k samples with the minimum prediction confidence level to be marked; (3) Delete these marked samples from the unlabeled data set U, and join them in the training data set L; (4) Retrain the classification model using the replaced training data set L; (5) Repeating steps (2)-(4) .

The test data set T is used to examine the prediction effect of the classification model produced currently. Once the performance increase of the new model on the test data set T is less than a given threshold, the marking will be stopped or the test data set T will be selected again [9].

B. KNN Classification Algorithm

K nearest neighbor (KNN) algorithm is the well-known non-parametric technique in the statistical pattern classification, owing to its simplicity, intuitiveness and effectiveness [10]. Its basic idea is to seek the most similar k samples in the feature space for those samples that will be classified, and to find out which category most of the k most similar samples belong to, and the category is just the one that those samples to be classified belong to. KNN algorithm only depends on recent samples to classify the category, so only a small amount of neighboring samples are considered when making the classification decision. So KNN algorithm is a suitable classificatory algorithm for those sample sets that their categories exit the intersection or overlapping.

The classification flow based on *KNN* algorithm can be divided into four phases for monitoring type of text: (1) Text preprocessing; (2) Feature selection; (3) Setting up text representation model; (4) KNN algorithm classification. Although it has no learning process, KNN algorithm still needs a great number of marked texts as the samples of training set. The calculation methods of text similarity degree are Euclidean distance, Chebyshev distance, Minkowski distance and so on [11]. The angel cosine is applied to calculate the similarity degree among texts in this paper, and the formula is as follows:

$$Sim(d_1,d_2)=\frac{\sum_{i=1}^{n}W_{1i}W_{2i}}{\sqrt{\sum_{i=1}^{n}W_{1i}^{\ 2}\sum_{i=1}^{n}W_{2i}^{\ 2}}} \qquad (1)$$

where W_{1i} and W_{2i} represent *i*th feature weights of vectors of texts d_1 and d_2 respectively.

If a part of *k* nearest neighboring documents belong to the same category, these documents and their categories should be graded by calculating the similarity sum of the same class documents and test documents. Because choosing documents is just to purely select *k* documents, the chosen documents may deviate, which requires put rewards and punishments of the category deviation into the score. So the score threshold should be set to determine the category of the test document. The score formula of KNN estimating test document *d* belonging to category C_i can be shown as:

$$Score(C_i, d) = \sum_{dj \in KNN} sim(d_j, d) y(C_i, d_j) - b_j \quad (2)$$

where

$$y(C_i, d_j) = \begin{cases} 1, d_j \in C_i \\ 0, d_j \notin C_i \end{cases};$$

d_j is the *j*th document among the selected *k* nearest neighbor test documents;

b_j is the optimal threshold of the category C_i obtained by calculation.

In short, KNN classification algorithm itself is simple, needn't learn, and has a low error rate. So KNN classification algorithm is the preferred classification algorithm in many cases that need classify texts.

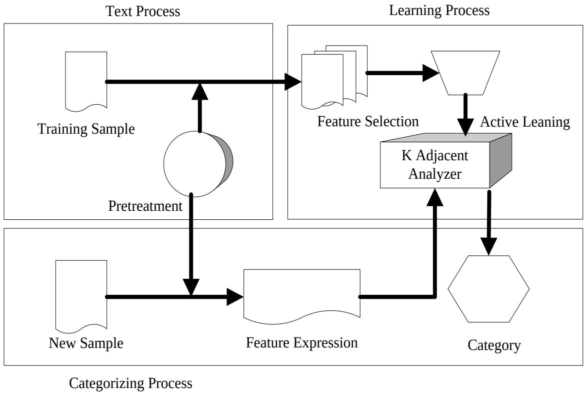

Figure 1. Classification process of classifier based on KNN algorithm

C. Question ClassifierBased on KNN Active Learning

For active questions classification learning method based on KNN, a small amount of texts need be marked by experts as classified standard documents of KNN classifier firstly. At the same time, unclassified documents are put into a learning system as selective materials, from which the boundary points will be selected. Similar to monitoring type of text learning and classification process, KNN classification process is as follows: (1) K nearest neighboring documents are chosen according to the text similarity, and the category score of unclassified samples can be obtained by the k nearest neighboring documents. As a result, the probabilities of unclassified samples belonging to various categories can be got; (2) The location entropy of samples can be got by information entropy formula based on the category probability obtained in (1). The sample will be placed into marked candidate pools if its location entropy reaches the threshold; (3) A random selective function is used to choose *n* samples which need to be tagged from marked candidate pools, the aim of which is to make the marked samples distributed more evenly on the boundary by means of random selection and to avoid the concentration of marked texts that will reduce the precision elevation of KNN classifier. Experts only need to mark the returned n texts that are just the boundary point texts of KNN classifier, thus the classifier precision elevation is realized, and the quantity of samples marked artificially is also reduced at the same time. There are two withdrawal conditions for active questions classification learning methods based on KNN, namely reaching a certain iterations or the accuracy of KNN classifier reaching the given threshold. The classification process of classifier based on KNN active learning is shown in Fig. 1.

III. Proposed Scheme

A. Questions Equivalent Standards and Definition

Equivalence relation is defined as an important binary relation in discrete mathematics. Assuming *R* is the relationship on a non-empty set *A*, if *R* is reflexive, symmetric and transitive, then *R* is called an equivalence relation on *A*. For any two elements of a set, namely *x* and *y*, if $< x, y > \in R$, then *x* is equivalent to *y*, which is denoted as $x \sim y$. Equivalence class of *x* on a set *A* is a set composed of all the elements of *A* that are equivalent to *x*, and is denoted as $[x]_R = \{y \mid y \in A \wedge xRy\}$, $[x]_R$ is the equivalence class of *x* about *R*.

The search engines usual recommend a collection of related problems to users after their questions are submitted in community question-answering system. These related problems are all ones put forward before by other users, and the questions in the correlative problem set are also related to each other. According to analysis, it can be got that the correlative relationship of questions is reflexive and symmetric. If all questions in the correlative problem set are correlative to each other, the correlative relationship of questions can be regarded as a kind of equivalent relationship. The related question set is an equivalence class of correlativity on question set *Q* [12]. So for research on large-scale question sets, the definitions of equivalence relationship and class are as follows:

Definition 1 (equivalence relationship: For a question set *Q*, if there exit variables Q_i and Q_j whose correlative degree is greater than a given threshold, then $Q_i \sim Q_j$, and is denoted by *RC*.

Definition 2 (equivalence class): For a question set *Q*, a set of variables *Q*' is called as an equivalence class if it meets the following conditions: (1) *Q*' is the subset of *Q*; (2) Any two variables on *Q*' meet the equivalence relation *R*; (3) There is no equivalence relation *R* between any variable Q_i on *Q*' and any variable Q_i on *Q*-*Q*'.

For example, Questions Q_1, Q_2, Q_3 and Q_4 are described as follows:

Q_1: "what is the social security contribution base in Changchun in 2011?"

Q_2: "What's the minimum social security contribution base in Changchun this year?"

Q_3: "what is the social security contribution base in Harbin in 2011?"

Q_4: "What's the social security contribution base, the upper limit and the lower limit in Beijing in 2011?"

The four questions all ask the size of the social security contribution base in an region in one year, and their structures and categories are very similar. Q_4 covers the first three questions' subjects and can be seen as their continuation. So the four questions and all related questions on the question set Q to the four questions can be seen as a large equivalence class.

B. The Determination of Questions Set Equivalence Relation

Correlativity between the questions is determined by relevant degree between questions. If their relevant degree is greater than the given threshold, the two questions are considered as correlative. The correlative degree of a question is a very fuzzy concept, and hasn't be defined definitely at present. The concept of correlative degree is given in this paper, that is, the judgment criterion of correlativity contains the following three determinant conditions: (1) The two questions belong to the same subject; (2) One of the two questions is a continuation or supplement of the another one. (3) The structures and categories belong to the same type. The first two determinant conditions are the broad understanding of the correlation. The third determination condition is a kind of empirical relationship that is established by a large number of observations at the community question-answering stations, and it is a powerful determinant condition for research on correlative questions of question-answering based on community [13]. Considering the above factors, the relevant degrees of questions are calculated respectively from three aspects, namely statistical information, semantic and subject information. The integral correlation degrees can be got by synthesizing the three aspects linearly.

Statistics information: based on the longest common word sequence

The longest public word sequence can be obtained by calculating the character strings of the longest word sequences after two question participles. If they include words that appear in the same sequence, the two questions are generally correlative. For Q_i and Q_j on set Q, their relevant degree can be formulated by:

$$\frac{|\mathrm{LCS}(\mathrm{Q}_i, Q_j)|}{C(Q_i \cup Q_j)} \tag{3}$$

where $|\mathrm{LCS}(\mathrm{Q}_i, Q_j)|$ is the vocabulary number of the longest public word sequence in Q_i and Q_j;

$C(Q_i \cup Q_j)$ is the sum of the vocabulary number of in Q_i and Q_j.

Semantic information

The semantic correlation between characteristic words can be obtained by synonymy thesaurus expanded version, based on which the relevant degree between questions can be got, for given w_1 and w_2, the semantic correlation degree can be formulated by [14]:

$$wordsim(w_1, w_2) = \frac{1}{distance(w_1, w_2)} \tag{4}$$

where $wordsim(w_1, w_2)$ is the shortest distance of w_1 and w_2 in semantic tree.

The question $Q_i(q_1, \cdots q_i, \cdots q_m)$ and the historical question $Q_j(d_1, \cdots d_j, \cdots d_m)$ constitute a $m*n$ dimensional matrix, in which each item is the similarity degree between words q_i and d_j.

$$\begin{bmatrix} wordsim(q_1, d_1)\cdots & wordsim(q_1, d_j)\cdots & wordsim(q_1, d_n) \\ \vdots & wordsim(q_i, d_j)\cdots & \vdots \\ wordsim(q_m, d_1)\cdots & wordsim(q_m, d_j)\cdots & wordsim(q_m, d_n) \end{bmatrix} \tag{5}$$

The similarity of questions can be formulated by:

$$semsim(Q, D) = (\sum\nolimits_{i=1}^{m} \frac{a_i}{m} + \sum\nolimits_{j=1}^{n} \frac{b_j}{n}) / 2 \tag{6}$$

where

$a_i = max(wordsim(q_i, d_1)\cdots wordsim(q_i, d_n))$

$b_j = max(wordsim(q_1, d_j)\cdots wordsim(q_m, d_j))$

Subject information: calculating the similarity degree using question topic information based LDA model.

LDA topic model is a kind of probability graph model, see Fig. 2.

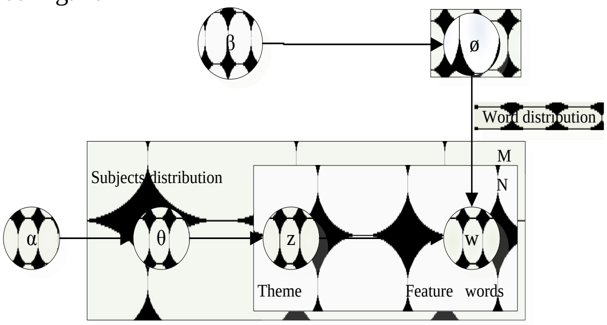

Figure 2. Probability graph model of LDA

LDA model describes the probability extracting process of words in document generated based on potential topics. The model is determined by parameters α and β. α reflects the relative strength among potential topics in document set, and β describes the probability distribution of all implied theme itself [15].

The process of generating documents is as follows: (1) For each topic t , a word multinomial distribution on t $\varnothing^{(t)}$ can be got based on Dirichlet (β); (2) For each question Q, a subject polynomial distribution on the document θ^Q can be got by Dirichlet (α); (3) For each word w_i in each question Q, a subject t is drawn from θ^Q, and a word is drawn from $\varnothing^{(t)}$ on the t as w_i

The probability of feature word w in question can be formulated by:

$$p_{1da}(w \mid Q, \hat{\theta}, \hat{\phi}) = \sum\nolimits_{z=1}^{T} p(w \mid z, \hat{\phi}) p(z \mid \hat{\theta}, d) \tag{7}$$

where

z is the subject corresponding to feature word w_i;

T is the number of topics;

$\hat{\theta}$ and $\hat{\phi}$ are priori estimates of θ and $\varnothing$.

Gibbs sampling extracting method can be used to extract the topic model based on LDA [12]. For a feature word w_i in a question Q, $w_i \in V = \{w_1, \cdots, w_V\}$, given a subject z_i=t, the probability of $\hat{\phi}_{w_i}^{z_i}$ and $\hat{\theta}^Q$ can be formulated by:

$$\begin{aligned}\hat{\phi}_{W_i}^{Z_i} &= \frac{n_{w_i,t}^{V_i,T}+\beta}{\sum_{i=1}^{V}(n_{W_i\cdot t}^{V_iT}+\beta)} \\ \hat{\theta}^Q &= \frac{n_{Q,t}^{D,T}+\alpha}{\sum_{j=1}^{T}(n_{Q,t}^{D,T}+\alpha)}\end{aligned} \tag{8}$$

For questions Q_i and Q_j, the V-dimensional word probability distribution based the subject z_i is:

$$\begin{aligned}P_{Q_i}^{z_i} &= (\hat{\phi}_{w_1}^{z_i},\cdots,\hat{\phi}_{w_{|Q_i|}}^{z_i},0,0,\cdots) \\ P_{Q_j}^{z_i} &= (\hat{\phi}_{w_1}^{z_i},\cdots,\hat{\phi}_{w_{|Q_j|}}^{z_i},0,0,\cdots)\end{aligned} \tag{9}$$

The questions Q_i and Q_j can be shown by subjects :

$$\begin{aligned}\vec{Q}_i &= (p_{Q_i}^{z_1},p_{Q_i}^{z_2},\cdots,p_{Q_i}^{z_T}) \\ \vec{Q}_j &= (p_{Q_j}^{z_1},p_{Q_j}^{z_2},\cdots,p_{Q_j}^{z_T})\end{aligned} \tag{10}$$

LDA is a probabilistic topic model. Each topic will appear in the document with a certain probability. The subject with small probability will produce ‘noise’ to the calclulation of relevant degree. So the subject can be chosen by a given threshold δ, that is, if $p_j^{z_i} < \delta$, then $p_j^{z_i} = 0 \quad (j = Q_i, Q_j; i = 1, 2, \cdots T)$. Finally the relevant degree between Q_i and Q_j can be calculated by using the most common cosine similarity.

$$cos(\vec{Q}_i,\vec{Q}_j) = \frac{\vec{Q}_i \times \vec{Q}_j}{|\vec{Q}_i| * |\vec{Q}_j|} \tag{11}$$

C. The Discription of Questions Set Equivalence Partition Method

After finishing the correlativity determination of questions, equivalence class partition of question set can be carried out. The steps of equivalence class partition are as follows: (1) Let $C = \{\{x\} \mid x \in Q\}$; (2) Find out all the equivalent pairs <x,y> to RC from Q; (3) A equivalent pair <x,y> is read in from the RC and the position of x and y in the C is judged , that is , x$\in$C$_i$ and y$\in$C$_j$ are got by looking up C$_i$$\in$C and C$_j$$\in$C; (4) If C$_i$$\neq$C$_j$, then C$_i$ is integrated into C, and C$_j$ is removed from C; (5) Turn to (2) until m equivalent pair s<x,y> are processed in R.

IV. Experiment

This paper will studyon the basic endowment insurance domain, and the experiment corpus in this paper is composed of 140000 real questions from Baidu and sina iAsk. Through comparing the classification results of our method with ones of the method based on rules and the method based on SVM, it can be indicated that the method proposed in this paper can obtain more accurate question classificatino result, with fewer labeled questions.

We will evaluate the accuracy of classification results, with the index value of Precision, Recall and F-Measure.

$$\text{Precision} = \frac{\text{Number of correct classification questions}}{\text{Number of questions to be classified}} \tag{12}$$

$$\text{Recall} = \frac{\text{Number of correct classification questions}}{\text{Number of questions in the right category}} \tag{13}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Presicion} \cdot \text{Recall}}{\text{Presicion} + \text{Recall}} \tag{14}$$

TABLE I. Corpus classification results

Categoriy of question	Set size
The insured	20,406
Payment	17,997
Personal account	23,241
Pension relationship transferring and continuing	37,677
treatment development and adjustment	34,327
Enterprise annuities	4,356
social management services.	1,985
Management	1,052

TABLE II. Corpus equivalence partitioning results

Categoriy of question	Divided block
The insured	725
Payment	679
Personal account	815
Pension relationship transferring and continuing	752
Treatment development and adjustment	839

The method proposed in this paper has two stages. Firstly, we categorize the questions by using the question classifier based on KNN active learning. Then the equivalence class partition is carried out for questions in each question category so as to realize more fine-grained questions organization. The results of every step are as follows:

Question classification experiment

By the reference to the book “basic pension insurance policy Q&A” published by the human resources and social insurance and social security ministry, the questions of pension insurance field are classified as eight subject categories, namely the insured, payment, personal accounts, pension relationship transferring and continuing, treatment development and adjustment, enterprise annuities, social management services and management. The number of questions in each category after classification is shown in Table 1.

The division of equivalence class

Because data sparseness will affect the LDA model, the hot problem categories are selected to classify such as the insured, payment, personal accounts, pension relationship transferring and continuing, treatment development and adjustment.

When determining the equivalence relation, the threshold is set to 0.5, namely if the relevance degree is greater than 0.5, there exits an equivalent relation between the two questions. The results of equivalent category classification are shown in Table 2.

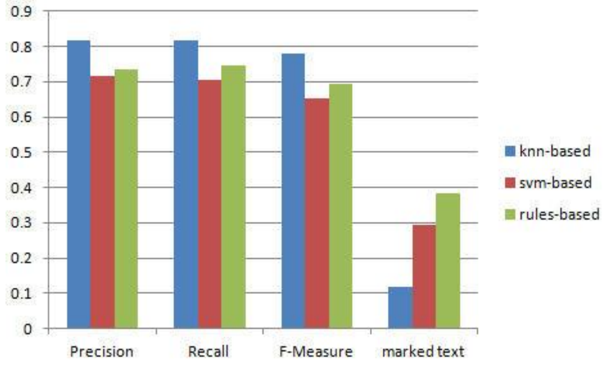

Figure 3. Comparison of index values in “the insured” category

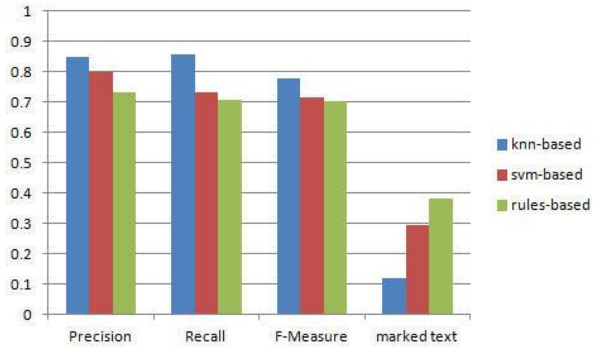

Figure 4. Comparison of index values in “Payment” category

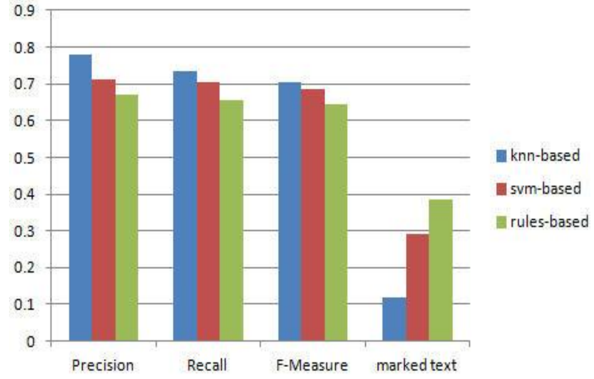

Figure 5. Comparison of index values in “personal account” category

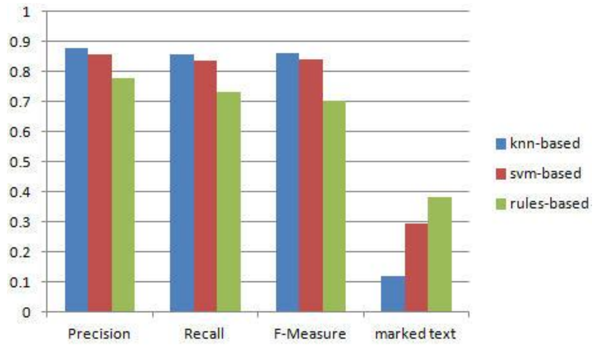

Figure 6. Comparison of index values in “pension relationship transferring and continuing” category

According to the equivalence partition experiment, it is found that although numbers of questions contained by these category sets are different, the number of divided blocks obtained by calculating is more evenly distributed. So it’s concluded that there are similarities among subjects contained by different category questions.

Next we will explain the validity of our method, i.e. it will obtain more accurate question classification result, with fewer labeled questions. As shown in Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, Fig. 9 and Fig. 10, the comparisons of classification results are indicated in the category of “the insured”, “payment”, “personal account”, “pension relationship transferring and continuing”, “treatment development and adjustment”, “enterprise annuities”, “social management services” and “management” respectively.

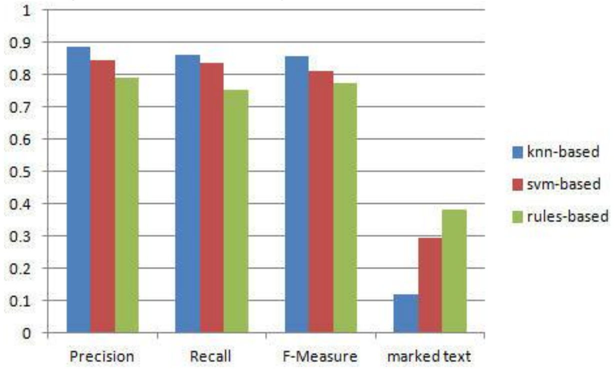

Figure 7. Comparison of index values in “treatment development and adjustment” category

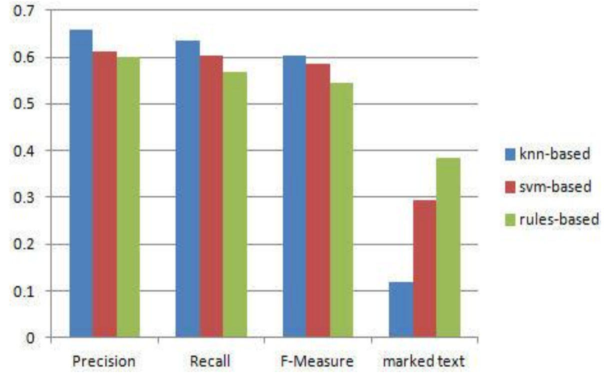

Figure 8. Comparison of index values in “enterprise annuities” category

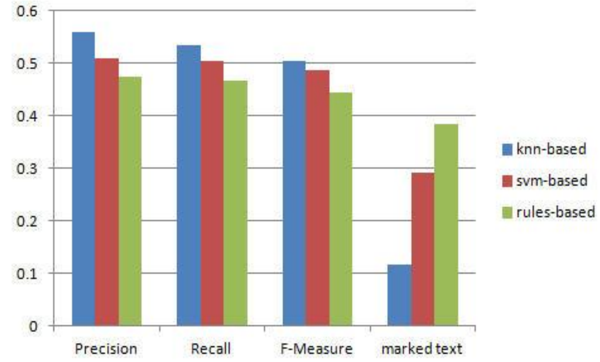

Figure 9. Comparison of index values in “social management services” category

Through analysing these results, we can conclude that the results of our method have higher accuracy, recall and F-Measure values.

V. Conclusion

Aiming at large-scale real set of questions from the Q & A communities, a new question classification algorithm is put forward based on KNN active learning to complete the question classification. At the same time, the equivalance class partition problem of question set is solved by integrating statistical information, semantic information and information. The pretreatment work of large-scale question set can be well completed based on the method.

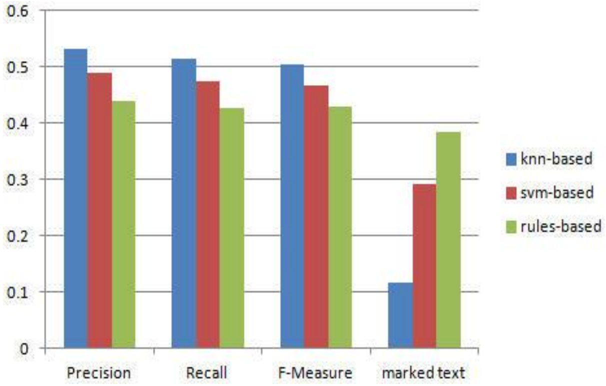

Figure 10. Comparison of index values in "management" category

Acknowledgment

This research was supposed by the "National Science & Technology Pillar Program" under the grant number of 2009BAH42B02 and 2012BAH08B02.

References

[1] Li Wei. "Question Classification Using Language Modeling", *Center of Intelligent Information Retrieval, Technical Report*, 2002.

[2] Li X, Roth D. "Learning question classifiers", *Proceedings of the 19th international conference on Computational linguistics, Association for Computational Linguistics*, 2002, 1 pp. 1-7.

[3] Zhang D, Lee W S. "Question classification using support vector machines", *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. ACM*, 2003 pp. 26-32.

[4] Carlson A, Cumby C, Rosen J, et al. "The SNoW learning architecture". *Technical report UIUCDCS*, 1999.

[5] http://ir.hit.edu.cn/phpwebsite/index.php?module=documents&JAS _DocumentManager_op=viewDocument&JAS_Document_id=131

[6] https://code.google.com/p/fudannlp/wiki/QuestionClassification

[7] S. Rajan, D. Yankov, S. J. Gaffney, et al. "A large-scale active learning system for topical categorization on the web", *Proceedings of the 19th international conference on World Wide Web*, 2010, pp. 791–800.

[8] R. M. Felder, R. Brent, "Active learning: An introduction", ASQ Higher Education Brief, 2009, 2(4) pp. 1–5.

[9] J. Attenberg and F. Provost, "Why label when you can search?: alternatives to active learning for applying human resources to build classification models under extreme class imbalance", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010 pp. 423–432.

[10] Jianping Gou, Taisong Xiong, Yin Kuang, "A Novel Weighted Voting for K-Nearest Neighbor Rule," *Journal of Computers*, Vol. 6, No 5, pp. 833-840, May 2011.

[11] Zhang zhuying, Huang Yulong, Wang Hanghu. "An efficient KNN classification algorithm". *Computer Science*, 2008 (03) pp. 170–172.

[12] Li Zhan, Chao Wenhan, Chen Xiaoming, et al, "The Chinese Community Q & A answer Quality Assessment and Prediction". *Computer Scienc*e", 2011, 38(6) pp. 230–236.

[13] Li Yuxiang, Li Shuanghong, Li Ru, "The issues related to community-based question-and-answer detection", S*ixth National Information Retrieval Conference* Proceedings, 2010.

[14] Tian Jiule, Zhao Wen, "Based on the the word of Cilin similarity calculation method", *Journal of Jilin University: Information Science*, 2010 (006) pp. 602–608.

[15] D. M.Blei, A. Y. Ng, M. I. Jordan. "Latent dirichlet allocation". *The Journal of machine learning researc*h, 2003, 3 pp. 993–1022.

[16] I. Porteous, D. Newman, A.Ihler, et al., "Fast collapsed gibbs sampling for latent dirichlet allocation", *KDD'08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008 pp. 569–577.

Qingsheng Wan, male, born in 14 February 1975, received the B.S degree in Agricultural Mechanization and Electrization from Northeast Agricultural University, Harbin, China, in 2000, and received the Master degree in Agricultural Mechanization and Electrization from Northeast Agricultural University, Harbin, China, in 2007. He is currently working towards the Ph. D in Harbin engineering University, Harbin, China. His research direction interests include artificial intelligence, natural language process and so on.