http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (5): 1061-1065, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1061.1065

Question Classification based on Rough Set Attributes and Value Reduction

[1]Li Peng and [2]Zhang Kai-Hui
[1]College of Computer Science and Technology,
[2]College of Economy and Management, Harbin University of Science and Technology,
Harbin 150001, China

Abstract: This study presents a method on automatic question classification through attribute and value reduction based on rough set theory. The core of the method is adopting statistical machine learning, with the assistance of a fair number of training corpus, attempts to automatically obtain useful and concise classification rules. Attributes reduction algorithm can omit the attributes which are unnecessary to decision classification in the decision table so as to simplify the decision table and increase the adaptability of decision process. The value reduction algorithm based on attributes significance can further eliminate the unnecessary information in the decision table. Comparing with the alternative means under the same data set and classification architecture, the experiment result is that the accuracy of the rough classification in this study is up to 86.20%, fine classification reaches 78.8%. It means that the method of this study is efficient.

Key words: Question classification, rough set, attributes reduction, value reduction

INTRODUCTION

Question Answering (QA) system is a hotspot of information processing research field in recent years and which is the development direction of the next generation search technology (Pasca, 2001). Generally, a complete QA system contains three main parts, including classification, information retrieval and answer extraction. Therefore, a rapid, effective, highly accurate of question classification technology is the basis of implementation of QA system, which will directly affect the accuracy of the final extraction answer (Kwok *et al*., 2001). The main task of question classification is to do the classification semantic category for the user's problem. The feature information contained in the question classification comes from user's question sentence, so the feature is less, which brings great difficulty in question classification. The classification system of question classification also needs changes according to the needs of the whole QA system, which makes the question classification, has high flexibility and needs good scalability and adaptability. For the research development of classification technology, it has the significantly practical significance to research the question classification method which has high accuracy and adaptability.

Question classification is based on the method of manual rules at first, the adaptability and extensibility of this classification method is very poor and time-consuming. In recent years, so the method based on statistical learning methods becomes the main way to solve this problem. At present, the statistical learning methods which have been applied on question classification are the Naive Bayes (Yu *et al*., 2005), Kernel Method (Suzuki *et al*., 2003), SnoW (Xin and Dan, 2002) and SVM (Dell and Sun, 2003) and so on. Those research results all show that statistical learning methods have good performance in question classification.

QUESTIONS ANALYSIS AND EXPRESSION OF DECISION TABLE KNOWLEDGE

Question classification system usually has two ways, namely parallel classification and hierarchical classification. Parallel classification system is a rough classification method; the classification method divides the question into several equal categories. If question classification adopts parallel classification, which tend to have relatively high accuracy of classification, but this classification system will increase technology difficulty for subsequent answer extraction part. Basing on rough classification, hierarchical classification system classifies each category for fine classification by further, the level increasing, the difficulty increasing. Hierarchical categorization is the trends of question classification research and also a difficulty of classification technology research.

Corresponding Author: Li Peng, College of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150001, China

The feature extraction is one of the most important parts in solving question classification by statistical machine learning method. Oriented objects of question classification are relatively short sentences, questions usually contain less word and therefore, containing information is relatively few. If wanting to improve the accuracy of classification, it needs as much as possible to extract useful information for classification from short sentences. In this study, each kind of character is expressed by a kind of attribute marks, for example, question word [Q-QW], keywords [Q-KW], named entity [Q-NE], basic noun phrases [Q-BNP] and so on. For each specific question, each kind of attribute corresponds with specific value. As the Fig. 1 represents some attributes of question feature extraction.

In particular, there is a default attributes NULL, which is come up with when a certain attribute value does not exist, it basically is to make the decision table as a complete information table, reducing the complexity of the processing.

Rule acquisition is to find useful and regular information from a large number of original data, which is to change the knowledge from an original form (original data form) to a new target form (form that is easy to process by human or computer, such as logical rules, etc.) (Guo-Yin, 2001). Discovery is based on the knowledge of rough sets theory, mainly with the help of information table, such a knowledge expression of the effective data sheets. Decision table is a special kind of information table, which can be used to generate decision-making rules and to solve question classification. The decision table is defined as follows:

Definition 1: A decision table is an information form knowledge expression system $S = \langle U, R, V, F\rangle$, here, U is an object set, also called the universe of discourse, $R = C \cup D$, is an attributes set, subset C and D are, respectively called condition attribute set and results attribute set, $D \neq \varnothing$. $V = \cup V_r$ is the collection of attribute values, V_r shows attribute domains of attribute $r \in R$. f: $U \times R \rightarrow V$ is an information function and allocate the attribute value of every object.

Through Table 1 the typical decision table illustration to further understand the composition of decision table and the role of obtaining the decision rules. The universe of discourse U actually is the sample collection of objects to be deal with; the condition attributes C is the classification attribute what can decide classification, namely characteristics; the results attribute D is the classification sets which classifies eventually, namely

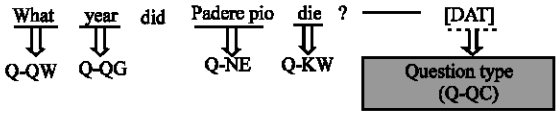

Fig. 1: Question feature extractions

Table 1: Decision information table of question samples

	C<condition attribute>					
U	C_1	C_2	C_3	……	C_M	D<decision attribute>
X_1	1	2	1	……	3	2
X_2	5	4	0	……	2	5
X_3	3	1	1	……	2	3
⋮	⋮	⋮	⋮	……	⋮	⋮
X_N	4	3	1	……	0	1

……: Means ellipsis

decision attribute. The training sample can be transformed into the decision table information expression. In order to meet the requirements of rough set method processing and program performing more convenient, some of the original attribute value has been quantified, replaced by discrete values. Decision attribute is also quantified for discrete number, how many types of classification, so many counts, once changed, only need to add or delete figures.

Every line in decision table is representative of a decision rule, in the following, some definitions of relevant decision rule are given, which show how to generate decision rule of from decision table.

Definition 2: Formula $A \rightarrow B$, its logical meaning called decision rule, A is called antecedent of rules, B is called consequent of rules, they express a causal relationship. Among them, the atomic formula of formula A only contains the condition attribute of decision table, atomic formula of formula B only contains the decision attribute of decision table.

As the decision table is shown in Table 1, the condition attributes set is $\{c_1, c_2\ c_3, ...\}$, decision attribute is, can generate the following decision rule:

$$(c_1, 1) \wedge (c_2, 2) \wedge (c_3, 1) \rightarrow (D, 2)$$

A: Rules antecdent B: Rules consequent

A sample of the decision table is representative of a decision rule, if all this decision rules enumerate out, it can get a decision rules set. However, such decision rules set only record one sample mechanically, do not get optimized and can not adapt to the changes in the situation. In order to extract rules which have big

adaptability from the decision table, it needs to reduce decision table, makes a record of the optimized and reduced decision table is representative of a characteristics of sample with the same rule, such decision rules have high adaptability.

REDUCTION ALGORITHM BASED ON ROUGH SET

Attribute reduction algorithm: The condition attributes in original decision table are not equally important; even some condition attributes are redundant. These redundant attributes, on one hand, is to waste resources; on the other hand, also interfere with people to make the right and concise decision. Therefore, the attribute reduction of the decision table, in the condition of keeping the classification capability of that condition attribute relative to the decision attribute, deletes the unnecessary attributes or not important attributes. Attribute reduction problem is an NP problem, there are some specific researches for attribute reduction (Guo-Yin and Hong, 2002), this study adopts attribute reduction algorithm based on feature selection, which is more classic and the procedures are as follows:

- Input decision table S = ⟨U, R, V, F⟩
- Output condition attributes set after the attributes reduction
- The first step calculating the related degree K (C, D) between condition attribute set C and decision attribute set D
- The second step REDU = C
- The third step while K (C, D) = K (REDU, D) do
  - Calculate the values of context CM of all attributes in REDU
  - According to the values of context, sort the attributes in REDU
  - Select the attribute a_j, a_j has the smallest value of context and K (REDU, D) = K (REDU\{a_j}, D)
- End while
- The forth step Output REDU

Value reduction algorithm based on the importance of attribute values: Attribute reduction of decision information table can omit unnecessary attribute of decision classification in decision table and realize reduction of decision table and what make for analyzing and discovering the attributes that contribute to decision classification, which improves the adaptability of the decision rules. However, to a certain extent, the attribute reduction removes redundant attributes of the decision, but which haven't fully removed redundant information of the decision table. Therefore, it puts forward value reduction question for further simplified decision table, which is based on the attribute reduction. In fact, the process of extracting rules in rough set theory is the process of value reduction for information table.

Analyzing the minimum value reduction, this study start from the value core, which is the biggest influencing attribute value for obtaining decision in the each record of information table, currently, some materials have introduced it. The value reduction method used in this study is a kind of value reduction algorithm based on the importance of attribute values, which has been accepted currently. Li-Yun *et al.* (1999) gives more detailed and special state in his study, combined with the merits of the above two kinds of reduction algorithm, this study proposes value reduction algorithm based on the importance of attribute values. Algorithm is as follows:

- Input: Information table S
- Output: Is value reduction of S

The first step Investigate line-by-line on condition attribute of each record in information table S, when deleting the column of the attribute:

- If generating the conflict records, retain the original attribute value of the conflict records, the value is irreducible
- If there is no conflict but repeating record, it will be labeled the attribute value of repeat records as "*", the value can be reduced
- If there is no conflict and repeating record, the attribute value will be labeled as "?", which shows whether or not to be reduced is undetermined

Attribute values which are not marked with "*" or "?" in the information table is the value core.

The second step, delete the repeat records which may generate and investigate every record containing label "?". If decision can be judged only by attribute values that were not marked, then do mark "*" instead of "?"; otherwise, modify mark "?" as the original attribute values; If the condition attributes of a record are all marked, modify the attribute items marked "?" as the original attribute values.

The third step, delete all of records that the condition attributes marked with "*" and possibly repeating records.

The forth step If only one condition attribute value of two records is different and the attribute is marked as

"*" in one of the records, then, for the record, if attribute value that not been marked can judge decision, delete another record; otherwise, delete this record.

In the new information tableafter reduction, all attribute values are the value core of the table, all records are the rules of the information table.

EXPERIMENTS AND PERFORMANCE ANALYSIS

Evaluation of comparative experiment: In order to verify and analyze the effect of the application on question classification by this method, it makes comparative experiments with the methods of Li Xin and Zhang Dell. In comparative experiments, it uses the same training data, test set and classification system. Training set includes 5,500 tagged question examples, test set is 500 problem examples evaluated by TREC l0 Q/A, classification uses two layers of the classification system, the first layer includes 6 rough classes, the second layer includes 50 fine classes.

Though this method, the accuracy of rough classification reached 86.20% and fine classification reached 78.80%, these results almost match with the best result of two kinds of methods which is contrast with, but in this two contrastive literature, researchers constantly investigate for the results of selecting various characteristics. The characteristic in this study is obtained by automatic optimization of attribute reduction, not existing the screening by artificial experimental. Dell and Sun (2003) also announced the classification results of the other statistical learning algorithm under the same experiment condition. Figure 2 and 3 display the results from other methods or different characteristics in the two articles, compared with those, it can be seen that the more excellent result obtained by the method in this study.

The result of attending TREC evaluation: We attended TREC2005 and TREC2006 two years' QA tasks evaluation, question classification is an important component of the system and its effects will directly affect the final result of the following answer extraction. In two years' evaluation, our question classification systems all adopt the single parallel classification system, such as Table 2 (Voorhees and Dang, 2005).

In the TREC2005 and TREC2006 QA evaluation, there are respectively 362 and 403 question types. The classification methods of these problems all adopt the strategy of the chapter. The number of each category of questions and the accuracy of classification was shown in Table 3.

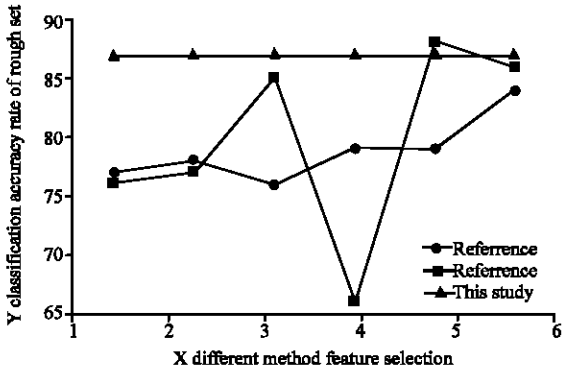

Fig. 2: Comparison of accuracy rate on rough set

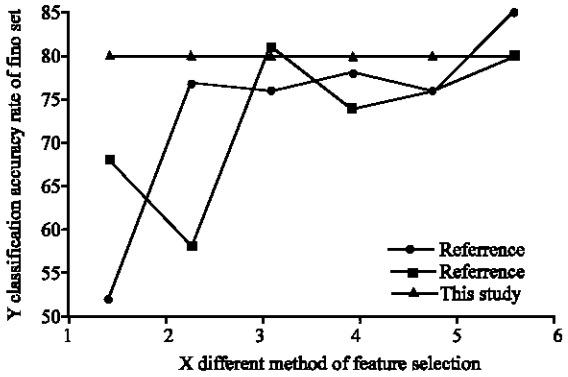

Fig. 3: Comparison of accuracy rate on fine set

Table 2: The question classification system of Insun QA system

Name	Place name	Organization	Time
Coin	Address	Period	Measurement
Number	Rate	Movie name	Name of name
Illness	Initialize	Color	Vocation
Brand	Tolephone	URL	Proper noun
General noun	Others		

Table 3: The result of question classification in TREC QA Track

	TREC 2005 QA track		TREC 2006 QA track	
FAQ type	No. of questions	Accuracy rate (%)	No. of questions	Accuracy rate (%)
Name	86	96.51	65	96.92
Place name	41	90.24	66	92.42
Organization	30	83.33	24	87.50
Time	56	96.43	81	97.53
Number	52	90.83	61	93.44
Others	97	80.41	106	79.25
Total	362	89.23	403	90.56

CONCLUSION

This study suggests the question classification method, through multiple knowledge acquisition procedure based on rough set theory supporting by (such as: data pretreatment, attribute reduction, value reduction

etc.), it realizes automatic generation and optimization of question classification rules, avoids a large number of labor on manual sorting rules and subjective interference of man-made selective characteristics, has the high accuracy of classification and degree of automation. Under the same test conditions, other methods of contrast experiments show that classification accuracy of rough classification and fine classification reached 86.20% and 78.80% in this method. The effect that obtained by the method applied in international TREC QA evaluating is good. Under the classification system in this study, the classification accuracy is achieved 92.54%, which has laid the solid foundation for QA system technique obtaining excellent achievement.

In the future research work, aiming at the research of question classification, it should do further research in the following respects. Firstly, feature selection should consider introducing the latest semantic information, such as block information, semantic role information; secondly, to strengthen question classification research on the multi-level classification system.

ACKNOWLEDGMENTS

Author would like to thank Doctoral Fund of Ministry of Education of China (No. 20102303120005) and Youth Research Fund of HUST, for funding this research.

REFERENCES

Dell, Z. and L.W. Sun, 2003. Question classification using support vector machines. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-Aug. 1, ACM Press, New York, pp: 26-32.

Guo-Yin, W., 2001. Rough Sets Theory and Knowledge Discovery. Xi'an Jiaotong University Press, China.

Guo-Yin, W. and Y. Hong, 2002. Decision table reduction based on conditional information entropy. Chinese J. Comput., 25: 759-766.

Kwok, C., O. Etzioni and D.S. Weld, 2001. Scaling question answering to the web. ACM Trans. Inform. Syst., 19: 242-262.

Li-Yun, C., W. Guo-Yin and W. Yu, 1999. Attribute reduction method and extraction rule based on rough set theory. J. Software, 10: 1206-1211.

Pasca, M.A., 2001. High-performance, open-domain question answering from large text collections. Ph.D. Thesis, University of Southern Methodist.

Suzuki, T.J., S. Yutaka and M. Eisaku, 2003. Question classification using HDAG kernel. Proceeding of the ACL Workshop on Multilingual Summarization and Question Answering, July 11, Sapporo, Japan, PP: 61-68.

Voorhees, E.M. and H.T. Dang, 2005. Overview of the TREC 2005 question answering track. Proceeding of the 14th Text Retrieval Conference (TREC'05), New York, pp: 1-15.

Xin, L. and R. Dan, 2002. Learning question classifier. Proceedings of the 19th International Conference on Computational Linguistics, (ICCL'02), Stroudsburg, PA, USA., pp: 556-562.

Yu, Z., L. Ting and W. Xu, 2005. Modified Bayesian model based question classification. J. Chinese Inform. Proce., 19: 100-105.