Study on the Application of the Theory of Rough Sets to Text Excavation

Dali Yin
ChangChun University of Science and Technology
Computer Science and Technology;
ChangChun ; China
Email: yindali2008@126.com

Yan Yan
ChangChun University of Science and Technology /
Computer Science and Technology;
ChangChun ; China
Email:yanyanxueer@126.com

***Abstract*—The rough set theory is a new calculation which deduces the concept categorization rules through attribute reduction with the categorization capacity remaining unchanged, and has wide application to data excavation and text excavation. This paper focuses on the analysis and improvement of the application of the rough set theory to text excavation, embracing the two aspects of attribute reduction based on clear matrix and correlation rules based on the Apriori algorithm. Compared with the classic algorithms, the improved algorithm can cut down a great deal on the expenditure of time and space in its operation, and improves considerably in its performance of processing the large-scale text database excavation. This research result has great theoretical significance to the application of the rough set theory and its algorithm to text excavation.**

Keywords-rough set; data excavation; text excavation; attribute reduction; correlation rules

Ⅰ. BACKGROUND OF RESEARCH

With the rapid development of the database technology and its application in many fields, the accumulated original data is on the sharp rise, and the present database system can realize the function of data saving, testing and so forth, but the present data excavation technology has some defects preventing it from discovering efficiently the relations and rules of data and constituting a practically valuable database. How to excavate valuable and potential knowledge in the mass of irregular data has formed the important research content of data excavation.

Text excavation is a key research trend in data excavation, which focuses on a large amount of text data, by means of quantitative calculation and qualitative analysis seeking for useful knowledge in irregular data. In the early of the eighties of the 20th century, Professor Z. Pawlak proposed the theory of rough sets [1], which is a new mathematical tool for dealing with incomplete information and inaccurate questions. The key idea of it is to find out categorization rules of concept through knowledge reduction with the categorization capacity remaining unchanged. The rough set theory has the features of needing no pre-experience, being capable of dealing with incomplete information and easy to be combined with other methods, therefore it is widely applied in the fields of data and text excavation. However, the research on the application of rough set theory to text excavation is now in its primary stage nationally, and many methods remain to be improved urgently. And what is more, it is a focal point in the present research.

Ⅱ. APPLICATION OF THE ROUGH SET THEORY TO TEXT EXCAVATION

Text excavation is a major branch of data excavation. With the swelling of text database, how to get useful knowledge through these data with high efficiency and high quality becomes a hot issue in the present research. Therefore, a large number of scholars and enterprises devote themselves in improving the method and efficiency of excavation algorithm. This paper will focus on the analysis and improvement concerning the two aspects of attribute reduction and correlation rules in the excavation in order to make text excavation more efficient. The attribute reduction and correlation rules are the classic algorithms in the rough set theory, whose division and relation are shown in Fig. 1.

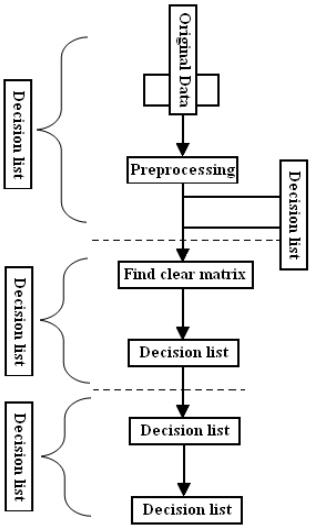

Fig.1. Division of Text Excavation Process

Attribute reduction and correlation rules are different parts of text excavation, but they are closely related in the process of excavation. First, attribute reduction is the precondition for the excavation of correlation rules. The attribute reduction will firstly reduce the attribute of decision list, cutting off the redundancy attribute and extracting a simple and plain reduced decision list, thus the quality of attribute reduction will directly affect the value of rules extracted from the correlation rules; secondly, the correlation rules are the intuitive

demonstration of the effect of attribute reduction. The excavation of correlation rules will extract correlation rules from the reduced decision list. If there are some defects in algorithm, it cannot make the right judgement of the effect of the algorithm of attribute reduction, and even cannot obtain right and useful correlation rules.

Ⅲ. RESEARCH AND IMPROVEMENT ON ALGORITHM

The extraction of attribute reduction and correlation rules constitutes the important part of the rough set theory, and in the meantime is an important step in text excavation. Due to the characteristics of multi-source and multi-variety of text data, the data stored in the knowledge base are full of both the good and the bad, therefore before the categorization of knowledge base, the complicated attribute set must be reduced, and the candidate data must be filtered and classified. As a result, on the one hand, the redundancy attribute can be removed reducing their interference with the classification of algorithm; on the other hand, the scale of knowledge is reduced, and the efficiency and quality of categorization of algorithm.

A. Attribute reduction algorithm based on clear matrix

1）The advancement of clear matrix and its thoughts

In the early 90s of the 20th century, A. Skowron advanced the method of representing knowledge with clear matrix based on the simplicity and plainness of attribute reduction by clear matrix and the capability of finding core attribute with high efficiency. Therefore, during rough set reduction, clear matrix has wide application, whose algorithm has the thoughts shown as follows

a) According to the definition, the clear matrix M of decision list is obtained;

b) For the non-void element C_{ij} of the clear matrix M, the corresponding extraction is yielded: $L_{ij} = \bigvee_{a_j \in C_{ij}} a$;

c) do conjunction to the extraction Lij, and the norm form of conjunction is yielded L: $L = \bigwedge_{C_j \neq 0} L_{ij}$;

d) For transformation of expression, transform the conjunction norm form L into the extraction norm form: L=∨Li;

e) Finally the output of the reduction result.

From the above clear matrix method with attribute reduction and combined with the literature[2][3] we can see that this method has the following defects:

a) The traditional attribute reduction algorithm based clear matrix starts from definitions and obtains clear matrix by means of the decision list. The time complexity of this algorithm is O(|C|2|U|2). Therefore, applying this algorithm to solving problems tends to need a mass of space and a long time. Especially in dealing with the large-scale data, problems like storage space insufficiency are likely to arise. Therefore, direct solution to the clear matrix of the decision list and then the realization of attribute reduction algorithm tends to result in unsatisfactory complexity of algorithm time and space.

b) From the course of attribute reduction with clear matrix we can see that the processing demands a transitional form to be yielded, which will result in waste in time and space. In addition, the logic expression of extraction yielded will have repeated terms, which will lead to the increase in calculation in the reduction of logical formula.

2） The Heuristic algorithm based on clear matrix

In order to raise the efficiency of algorithm, aiming at the deficiency of the traditional attribute reduction algorithm based on clear matrix, the algorithm is improved in the following ways:

a) This paper advances that the decision list should be simplified first, whose procedure is realized by quick solution to the algorithm of the division of U/C, and whose time complexity is O(|C||U|)， and then the simplified clear matrix and the attribute reduction based on the simplified clear matrix are found after the decision list is simplified, and the equivalent value is proved between attribute reduction based simplified clear matrix and attribute reduction based on the original clear matrix.

- Algorithm 1： demands the input: the decision list $S=(U,C,D,V,f)$， $U=\{x_1,x_2,\cdots,x_n\}$， $C=\{c_1,c_2,\cdots,c_n\}$；
  The output: U'_{pos}, U'_{neg}, m_i, M_i, M_D, M_D(i=1,2, ⋯,k).

① Find the statistics of every element and decision attribute D in conditional attribute sets, and record the maximum and minimum of $f(x_j,c_i)$ and $f(x_j,D)(j=1,2,\cdots,n)$ with Mi , mi and MD ,mD;

② put the element $x_1,x_2,\cdots,x_n$ in the domain U in the linked list L one by one, with the pointer pointing to x_1;

③ For（i = 1； i < k+1; i++）

i Construct the void queue of Mi-mi+1, and put the element x in the linked list L in the corresponding queue of $f(x,c_i)-m_i$ in turn;

ii. Reconstruct the linked list L′, revise the head and tail pointers of the linked list L, enabling them to point to the first and the last non-void queue respectively, and consequently rebuild queues amounting to Mi-mi+1 into a new linked list L′;

④ suppose the element sequence in the new linked list L′are $x'_1,x'_2,\cdots,x'_n$； t = 1； $B_t=\{x'_1\}$;

For(j=2； j < n+1； j++)

If any element for the conditional attribute C can fulfill $f(x'_j,c_i)= f(x'_{j-1},c_i)$, then $B_t=B_t \cup \{x'_j\}$; otherwise {t = t+1； $B_t=\{x'_j\}$ }；

⑤ let $U'_{pos}=\emptyset$， $U'_{neg}=\emptyset$，

For(i =1； i < t+1； i++)

If elements contained in B_i have equal value in decision attribute, then extract the first element in B_i and incorporate it intoU'_{pos}, otherwise incorporate it into U'_{neg}

b) From the definition of clear matrix we can see that if a certain object in matrix only contains one attribute, then this attribute is the necessary attribute that

distinguishes it from other objects, that is, the sets of objects of independent attribute contained in clear matrix are the relative key attributes in the decision list. Therefore, in the solution to the clear matrix in the decision list and the output of attribute reduction result, the objects that contains key attributes can be taken out first, and the calculation be continued on the simplified matrix, which raises the efficiency of calculation and in the end attaches key attributes to each simplified expression.

- Algorithm 2 ： implification of attribute reduction process

① Based on the decision list S construct the clear matrix M, and extract the objects containing only individual attributes, which constitutes the core of S CORE(C). If $CORE(C) \neq \Phi$, we draw $R \cup CORE(C) \rightarrow R \neq \phi$；

② Obtain all objects consisting only of key attributes in T, that is:

$Q = \{B_i \mid B_i \cap CORE(C) \neq \phi, B_i \in T\}$

And the simplified decision list is : $S - Q \rightarrow S$；

③ Transform the attribute combination in the simplified decision list S with the conjunction norm form;

④ Transform P with the extraction norm form;

⑤ Output of reduction results.

From algorithm improvement we can see that the first operated simplification of decision list enables it to have the time and space complexity reduced to $O(|C||U|)$, and consequently reduces the operation scale of the yielded clear matrix, saving considerably in computer expenditure; the extraction of sets with key attributes from clear matrix reduces the number of attributes involved in calculation, and greatly raises the efficiency of algorithm.

B．Excavation of correlation rules based on the Apriori algorithm

1）Advancement of the Apriori algorithm

In 1993, Agrawal and other people first advanced the question of discovering correlation rules between customer transaction database sets based on customer transactions, as well as the Apriori algorithm based on frequent sets, which is one of the algorithms mainly used in the excavation of correlation rules at the present time.

The Apriori algorithm[5] is the most efficient algorithm in excavation of the sets of frequent terms in the Bull correlation rules at present, whose core is the recursion algorithm based on the frequent set idea in two stages. This algorithm first finds all frequent sets, and then searches each tier in turn. When it searches for the k time, a frequent candidate set is formed in the next tier. By scanning database, a complete frequent set of item (k+1) is got. By recurring， all frequent sets are obtained. It is mainly achieved by the following two steps:

a） Linking. First in the affair database find the set of term which is not less than the minimum supporting degree， and constitute the frequent set of term. Suppose l_1 and l_2 are the two set of term in the frequent sets, if L_1 and L_2 fulfil:

$(l_1[1]=l_2[1]) \wedge (l_1[2]=l_2[2]) \wedge \cdots \wedge (l_1[k-2]=l_2[k-2] \wedge (l_1[k-1]=l_2[k-1])$

Then link l_1, l_2 to the result set of term $l_1[1]l_2[2]\cdots l_1[k-1]l_2[k-1]$ .Recorded as C_k among which $l_i[j]$ represent the j item of l_i.

*b）*Deletion. Scan the linked set of term C_k which is obtained in the first step to define the count of credibility of each candidate set of term. As the scale of C_k may be enormously huge, the efficiency of searching each tier should be raised by the quality of the Apriori in the counting which is that “all non-void subsets in the frequent set of term must be frequent”. That is to say, if the subset of term (K-1) in a certain K set of term is not in L_{k-1}, then this candidate set of term is not frequent. Therefore, it can be removed from C_k.

2）Analysis of the defects in the Apriori algorithm

a) Frequent scanning of database. From the realization of the Apriori algorithm, the count of each set of term in a candidate set can be performed only by one scanning of affair database. Suppose the number of set of term in the candidate set C_k is $|C_k|$, the count contained in the affair database D is n, and the size of each record is p, then the time for calculating candidate set C_k is $O(|C_k| np)$, and consequently it can be found that the time for calculating all candidate sets is $O(\sum_k |C_k| np)$.

Obviously, the time for repeated scanning of database in calculating candidate sets is proportional to the volume of D and the number of candidate sets, therefore when there is a huge mass of the count of database, algorithm will consume a great deal of time and even cannot be

performed.

b) When the deletion step is performed in the algorithm, for any c $\in C_k$,it must be judged whether the set of term (k-1) as many as k belongs to L_{K-1}, and if one is found that does not belong will be eliminated by c. As L_{K-1}, needs to be scanned repeatedly, it will seriously affect the efficiency of algorithm when the scale of C_k is larger[6].

3）Key idea of the improvement of the Apriori algorithm

The realization of the Apriori algorithm needs repeated scanning of affair database, therefore the scale of database will directly affect the efficiency of the algorithm. Trying every means to reduce the scale of database is the important measure for raising the efficiency of algorithm.

The core idea of the improvement of algorithm is that when the frequent set of term at the k time's scanning is formed, with the candidate set of term of the k dimension forming by the k-1 set of term, and on the basis of the quality of the Apriori algorithm, any element contained in the frequent set of term of k dimension will be will be counted as k-1, that is to say, if the number of element in the frequent set of term of k-1 is less than k-1, this term can be deleted, so that a large amount of combinations that may combined with this term can be removed. Then all the combinations of the k-1 dimension in the frequent set of term of k are checked to see whether each combination is contained in the frequent set of term of k-1. If not contained, the combination will be deleted. After the cyclical processing in each step, the frequent set of term of k is got, scanning each affair in the affair database. If the affair does not contain any subset of the set of term of C_k, this affair will be exchanged for the affair at the end of database, and be deleted with mark. After the scanning of database, the new affair database reduced by recorded number can be obtained. The scale of database is reduced considerably, which makes it convenient to excavate data with high dimension of the recording, saving considerably in the expenditure of I/O.

Ⅳ. CONCLUSION

This paper focuses on the algorithm of attribute reduction based on clear matrix and the excavation of correlation rules based on the Apriori algorithm. By the simplification of the original decision list, which first removes its redundancy attribute and results in simplified clear matrix, it proves that the simplified clear matrix and the original clear matrix have equal value, and that after the extraction of core attributes, high-quality attribute reduction can be got with high efficiency, which provides the concise and high-quality reduced decision list for the next extraction of correlation rules; the traditional Apriori algorithm needs repeated scanning of database resulting in waste in the resources of the system. The Apriori algorithm has the quality that "all the subsets in the frequent set of term must be frequent", on whose basis the term which does not fulfil this quality is deleted while the candidate sets of term are formed by the decision list, so that the scale of database is greatly reduced and the efficiency of algorithm is raised, with the correctness of the extraction of correlation rules ensured. The experiment shows that with the improvement of the two algorithms, the efficiency of excavation is greatly raised in the overall excavation of text, the time complexity of algorithm reduced, which provides a reliable guarantee for the stability of system, so that more reliable excavation results can be obtained.

REFERENCES

[1]. Wang Yu. Study on the Rough Set Theory and its Application a doctoral thesis of the Electronic Technology of Xi'an. 2006

[2] Wang J, Wang R, Miao D Q, et al .Data enriching based on Rough Sets theory. Chin J Comput,2005, 21(5):393

[3] Wang J, Reduction algorithms based on discernibility matrix: the ordered attributes method. J Comput SciTechnol,2001,16(6):489

[4] Xu Z Y. Liu Z P. Yang B R. et al. A quick attribute reduction algorithm with complexity of max(O(|C||U|)，O(|C|2|U/C|)).Chin J Comput . 2006, 29(3) :391

[5] Qu Chunjin. The Apriori-TIDS algorithm Design and its Application in the Excavation of Information in Educational Decisions. July, 2005.

[6] Sheng Li; Liu Xiyu; Gao Ming. Study on the Algorithm of Data Excavation based on the Rough Set Theory. Shandong Sciences. 2005.

[7] Song Jing; Lu li. Categorized Excavation based on the Apriori Algorithm. Journal of Xihua University. 2007