A Clustering Algorithm of E-learningers Based on Rough Set

Yunhua Wang[1,2]
1. School of Computer, Wuhan University
2. School of Computer Science and Technology,
Wuhan University of Technology
Wuhan, China
E-mail: yhwang@whut.edu.cn

Huiyan Ke
Soil and Water Conservation Monitoring Centre
Department of Water Resources of Hubei Province
Wuhan, China
E-mail: huiyankeke@sina.com

***Abstract*—The demand for individualized teaching from E-learning websites is rapidly increasing due to the huge differences existed among E-learning learners. A method for clustering E-learningers based on rough set is proposed. The basic idea of the method is to reduce the learning attributes prior to clustering, and therefore the clustering of E-learningers is carried out in a relative low-dimensional space. Using this method, the E-learning websites can arrange corresponding teaching content for different clusters of learners so that the learners' individual requirements can be more satisfied.**

Keywords- concep; E-learning; clustering algorithm; rough set; individualized teaching

I. INTRODUCTION

Modern distance education is a new type of education form based on Web. The learners in network differ in many ways, such as knowledge level, taste of media form, learning obstacle, etc. But presently E-learning websites generally provide the learners with singular teaching mode, which can not meet the learners' personal needs. So the key to solving this contradiction should be converting "regarding websites as the center" into "regarding learners as the center", i.e., the teaching websites should pay more attention to the individual characteristics of learners during their teaching course so that the websites can teach the learners in accordance with their aptitudes.

Clustering is one of the most important and useful techniques that contribute to the individualized service[1]. Researchers, both home and abroad, have proposed many methods for clustering Web users in recent years[2-4]. However, there are commonly two disadvantages in those methods. Firstly, clustering is only based on certain information of users' accessing logs, which can not characterize the users completely, e.g., only the time that the users stayed on Web pages is taken into consideration in Ref.[4]. Secondly, there are few effective measures to deal with noise data which results in redundant elements of the user attribute vector. Among various clustering algorithms [5], *k*-means may be the most well known and commonly used method, but it is very sensitive to the irrelevant elements of attribute vector, which mislead the process of clustering[6]. The rough set theory introduced by Pawlak Z in 1982 is propitious to deal with redundant attributes.

In this paper, on the analysis of learning attributes of E-learningers, a model for clustering E-learningers and the corresponding algorithms are proposed, which use the reduction method of rough set to deal with redundant attributes. It can not only clean noise data, but also improve the efficiency of clustering algorithm. So it can be applied to the individualized teaching in distance education.

II. CLUSTERING MODEL BASED ON ROUGH SET

2.1 Some basic concepts of rough set

For the convenience of description, we introduce some basic notions of rough set at first.

Definition 1 An information system (IS) is defined as $S=(U,A)$, where U is a non-empty, finite set of objects, called the universe, and A denote a non-empty, finite set ofattributes. With each attribute $a\in A$, we associate a set V_a, of its values, called the domain of a.

Definition 2 For the information system $S=(U,A)$, in this paper, $U=\{u_1,u_2,\cdots,u_n\}$ denotes the set of Web learners, and $A=\{a_1,a_2,\cdots,a_p\}$ denotes the set of learning attributes.The discernibilty matrix of IS is a symmetric $n\times n$ matrixM, the element c_{ij} of which is defined as $c_{ij}=\{a\in A\mid a(u_i)\neq a(u_j)\wedge u_i,u_j\in U\wedge i,j=1,2,\cdots,n\}$, where $a(u_i)$ denotes the value of a ttributeafor object u_i.

The element c_{ij} consists of the set of attributes in which objects u_i and u_j differ, so the discernibilty matrix includes the distinguishing information for all pairs of objects. It can be used to discover the core attributes and attributes reduction.

Definition 3 Given an IS $S=(U,A)$, any subset B of A determines an indiscernibility relation $INB(B)$, which is defined as $IND(B)=\{(u,u')\in U^2\mid\forall b\in B,b(u)=b(u')\}$

Published by Atlantis Press, Paris, France.

A reduction of A is a minimal set of attributes $B \subseteq A$ such that $IND(A) = IND(B)$ and there is no such $b \in B$ that can satisfy the equation $IND(B) = IND(B - b)$.

2.2 Clustering Model

To implement individualized teaching, the teaching website should cluster the huge number of learners into classes appropriately at first so that learners within one class are similar to each other in their learning attributes, but are very dissimilar to learners in other classes. Thus the website can arrange the teaching strategy and content for each class correspondingly. In k-means algorithm, objects are described byattribute values, and clustering is based on the similarity between the objects, which is computed based on the distance between each pair of objects. As to this paper, the similarity between each pairof learners is calculated based on a learner-attribute matrix, the columns ofwhich are labeled by a set of attributes and the rows of which are labeled by E-learningers.Considering the reduction method of rough set can remove redundant attributes effectively without weakening the distinguishing ability of information table, we give the model for clustering of E-learningers based on rough set, as Figure 1 shows.

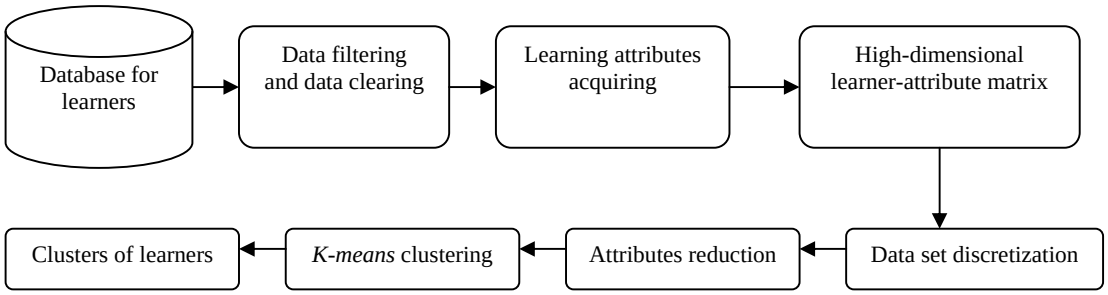

Figure 1. The model of clustering of E-learningers

2.3 Learning Attributes Selection

The E-learningers'various favors often reveal in course of their learning, e.g., some like to learn by cartoons and pictures while some others prefer to being guided by certain number of examples. If we ignore these diversities, the enthusiasm and interest of learners would disappear gradually under singular teaching strategy and teaching mode. The research of Ref.[7] shows that the Web learners'studying effects are affected by many factors, such as the attitude toward E-learning,the learning style, the desire for studying judgment, the time-spending on network, and the activities on Web, etc.

In the norm of Distance Learning Technology Standards (DLTS) of china, it describes that the data forming the learner mode comes from 7 concrete items, i.e., learning record, effect information, work collection, school report, personal information, trending information and learning style. For this reason, we choose learning attributes from three sources. The first one is the registered information, such as sex, age, academic degree, media taste, attitude to E-learning, learning purpose, understanding degree of educational technical resource (computer, network and multimedia technology), and rudimental knowledge about selected courses and so on. The second one is the record of all the grades of finished courses, which can predict how well a learner would achieve in his or her successive courses and the third one is the statistic information comingfrom Web logs, such as learning paths, ways of seeking help, feedback information to website, information of using White Board and BBS, etc., which can reflect the enthusiasm and initiative of learners as well as their adaptability to network.

To facilitate the application of rough set, we map the value set of each selected attribute to a successive integer sequence, e.g., the mapping relations for time-spended on BBS are showed as TABLE 1.

TABLE I. THE DISCRETE VALUES FOR THE TIME-SPENDED ON BBS

Time of using BBS	Attribute value
[0, T/4)	0
[T/4, T/2)	1
[T/2, 3/4T)	2
[3/4T,T]	3

III. ALGORITHM REALIZATION

3.1 Reduction of Learning Attribute

The gained learning attribute vector finally may include irrelevant elements because of the abroad source of the selected attributes. By the most popular distance measure adopted in k-means, which is defined as

$$d(i, j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \qquad (1)$$

where $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $X_j = (x_{j1}, x_{j2}, \cdots, x_{jp})$ are two p dimensional data objects described by attributes, it is obvious that such distance involves all the attributes. But in general, it is a part of the whole attributes that is most relevant to clustering, and the others, called unimportant or redundant attributes, would mislead the process of clustering and reduce the qualityof clustering, so it is necessary to deal with the redundant attributes prior to clustering.

Attributes reduction, a core problem of the rough set research community, can select out the indispensable attributes from the whole attributes. Compared with the whole attributes, the selected part has the same (or almost the same) ability to categorize the objects, so we can use the selected part instead of the whole in our clustering. The discernibilty matrix, introduced by Skowron A, can convert the process ofattributes reduction to the process of transforming conjunctive normal form to disjunctive normal form, the main idea of which is to use logic mathematics to make each intersection set, the reduced attribute set intersects with each elements of the discernibilty matrix, non-empty, so each pair of objects have at least one attribute to distinguish. If an element of the matrix only includes single attribute, called core attribute, then it is the only attribute that can distinguish the two corresponding objects inthematrix. Core attribute is necessary, so the core attributes can be the starting set, and other useful attributes are hidden in those matrix elements that don't include any core attribute. We give the reduction algorithm as follows:

Step 1 Compute the discernibilty matrix M for the information system of E-learningers;

Step 2 For each $|c_{ij}| = 1$ element of matrix, choose the attributes included in it to formthe core attribute set C_0;

Step 3 Establish the conjunctive normal form $L = \bigwedge_{c_{ij} \neq \varnothing, C_0 \cap c_{ij} = \varnothing} c_{ij}$, where c_{ij} is represented as a disjunctive normal formof its attributes.

Step 4 Convert the conjunctive normal form L to adisjunctive normal form $L' = \bigvee_i L_i$ .The attributes of each L_i combined with the C_0 turn out a result of reduction.

This method removes the redundant attributes based on the attribute values of the objects in universe, independent with any people's prior knowledge, so it is more objective and credible.

3.2 Algorithm for Clustering of Learners

It is obvious that attributes reduction reduces the dimensions of the attribute vector on the premise of preserving the essential attributes. Supposing after attributes reduction, we get an $n \times p$ dimensional learner-attribute matrix, the number of clusters k, the max iterative time $T_{\max}$, and the error threshold θ. The clustering algorithm proceeds as follows:

Step 1 Normalize the data set in the learners'attribute matrix so that each attribute has equal weight. This can be achieved by replacing x_{ik} with z_{ik} by

$$z_{ik} = \frac{x_{ik} - \min\limits_{1 \le j \le n}\{x_{jk}\}}{\max\limits_{1 \le j \le n}\{x_{jk}\} - \min\limits_{1 \le j \le n}\{x_{jk}\}} \qquad (2)$$

where x_{ik} denotes the value of the ith learner on the kth attribute;

Step 2 Arbitrarily chooseklearners as the initial cluster centers $c_1, c_2, \cdots, c_k$. The formula for computing the centers is as following

$$\overrightarrow{c_j} = \frac{1}{|C_j|} \sum_{x_i \in C_j} \overrightarrow{x_i} \qquad (3)$$

where $|C_j|$ denotes the number of learners of class C_j;

Step 3 Compute the distance between each learner and each class center by Eq.(1), and assign each learner to the class to which the learner is the closest, based on the center of the class;

Step 4 Update the class centers, i.e., calculate the center for each class by Eq.(3);

Step 5 Compute the error functionwhich is defined as

$$E = \sum_{i=1}^{k} \sum_{x_i \in C_i} |x - c_i|^2 \qquad (4)$$

If $E < \theta$, then end; otherwise, if $t > T_{\max}$ (t is the time that the algorithm already has executed), then end; otherwise, empty all the objects for each class, and go back to Step 3.

IV. EXPERIMENT ANALYSIS

Suppose that Table 2 is a discrete learner-attribute matrix, the attribute set $A = \{a_1, a_2, \cdots, a_7\}$ of which is only a small potion of the whole for simpleness. We get TABLE 3, a reduction of TABLE 2, by computing out a reduced set $B = \{a_2, a_3, a_5, a_7\}$ of A using the attributes reduction algorithm introduced above. It is obvious that the discernibilty ability of TABLE 3 is equal to that of TABLE 2 for the 9 learners, thus the clusteringon high-dimension data has been converted to the clustering on low-dimension data.We realize the *k*-means clustering of the learners in TABLE 2 and TABLE 3 respectively using VC 6.0 on the plat form of Widows XP. To execute algorithm 10 times repeatedly, the average iterative times for TABLE 2 is 2.7, and that of TABLE 3 is 2.2. The result shows that both the space and the time of the latter algorithmhave reduced because of attributes reduction. We believe that the effect of attributes reduction will be much greater as the number of the learners' amount to thousands upon thousands.

TABLE II. A SMALL LEARNER-ATTRIBUTE MATRIX

	a1	a2	a3	a4	a5	a6	a7
u1	0	3	4	0	1	5	0
u2	0	3	4	0	1	5	2
u3	0	3	2	0	1	2	2
u4	0	1	1	3	3	5	0
u5	1	2	4	0	1	5	0
u6	1	3	4	0	2	5	0
u7	0	2	3	1	1	2	2
u8	0	3	3	2	1	3	2
u9	1	1	3	1	2	2	2

TABLE III. A REDUCTION OF TABLE 2

	a2	a3	a5	a7
u1	3	4	1	0
u2	3	4	1	2
u3	3	2	1	2
u4	1	1	3	0
u5	2	4	1	0
u6	3	4	2	0
u7	2	3	1	2
u8	3	3	1	2
u9	1	3	2	2

V. CONCLUSIONS

How to improve individualized teaching is one of the most important and difficult issue in the developing of advanced distance education. In this paper, we proposed a model for clustering E-learningers based on rough set theory, the corresponding reduction algorithm and clustering algorithm. This method can provide service for teaching learners in accordance with their learning attributes. Further research work is to mine and apply the dynamic individual knowledge of Web learners in order to serve for the demand of real-time individualized teaching.

REFERENCES

[1] Y. Liu, S.S. Ge, C. Li, Z. You, “k-NS: A classifier by the distance to the nearest subspace”, IEEE Transactions on Neural Networks, vol.22, pp. 1256–1268, 2001.

[2] Y. Liu, Z. You, L. Cao, “A novel and quick SVM-based multi-class classifier”, Pattern Recognition, vol.39, pp. 2258–2264, 2006.

[3] J.W. Han, M. Kambr, Data Mining-Concepts and Techniques, Beijing: Higher Education Press, 2001.

[4] H.Y. Wang, Y.G. Zhang, “An improved artificial fish swarm algorithm of solving clustering analysis problem”, Computer technology and development.vol.20, pp.84–91, 2010.

[5] Z.Pawlak, RoughSet-TheoreticalAspects of Reasoning about Data Dordrecht, Boston: Kulwer Academic Publishers, 1991.

[6] S.B. Gong, Y.C. Guo, “Genetic algorithm based on fuzzy cluster analysis”, Fuzzy systems and mathematics. Vol.24, pp.123-128, 2010.

[7] L. Wang, W.M. Liu, H. Yang. “What Student’s Characteristics Affect the Efficiency on E-Learning”, Proceedings of Global Chiese Conference on Computers in Education, Beijing: China Central Radio and TVUniversity Press, Dec.2002, pp.27-30.