Research Article

A Coupled User Clustering Algorithm Based on Mixed Data for Web-Based Learning Systems

Ke Niu,¹ Zhendong Niu,¹ Yan Su,¹ Can Wang,² Hao Lu,¹ and Jian Guan¹

¹*School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*

²*Digital Productivity, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, TAS 7005, Australia*

Correspondence should be addressed to Zhendong Niu; zniui@bit.edu.cn

Received 24 April 2015; Accepted 15 June 2015

Academic Editor: Francisco Alhama

Copyright © 2015 Ke Niu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In traditional Web-based learning systems, due to insufficient learning behaviors analysis and personalized study guides, a few user clustering algorithms are introduced. While analyzing the behaviors with these algorithms, researchers generally focus on continuous data but easily neglect discrete data, each of which is generated from online learning actions. Moreover, there are implicit coupled interactions among the data but are frequently ignored in the introduced algorithms. Therefore, a mass of significant information which can positively affect clustering accuracy is neglected. To solve the above issues, we proposed a coupled user clustering algorithm for Web-based learning systems by taking into account both discrete and continuous data, as well as intracoupled and intercoupled interactions of the data. The experiment result in this paper demonstrates the outperformance of the proposed algorithm.

1. Introduction

Information technology and data mining have brought great changes to education field. Web-based learning is a significant and advanced type of education, which utilizes computer network technology, multimedia digital technology, database technology, and other modern information technologies to learn in digital environment.

At present, many education institutions and researchers commence the study of Web-based learning systems. They mainly study the systems' composition, the construction of a learning mode, the design and development of hardware, relevant supportive policies and services, and so forth. Meanwhile, an increasing number of Web-based learning systems develop rapidly, for instance, online study communities and virtual schools [1]. MOOCs (Massive Open Online Courses) are open online study platforms which provide free courses to students. It was initiated by America's top universities in 2012 and had a participation of more than 6 million students from around 220 countries, within a year [2]. In these systems, all learners received same learning resources but no customized or personalized learning services. They are short of analysis

on learners' behaviors and individual features; thus, scientific guidance and help is necessarily needed. In addition, there is a mass of learning resources in the systems, which leads to a big challenge, how to tease out the most wanted and suitable resource.

User clustering can dig out hidden information from a large amount of data. By clustering users in different ways, Web-based learning systems can provide personalized learning guides and learning resources recommendation to learners. This can greatly improve learning efficiency in these systems.

Recently, there have been some cases of applying user clustering algorithms in Web-based learning systems. In order to choose suitable learning method, clustering was addressed [3]. Lin et al. proposed the kernel intuitionistic fuzzy *c*-means clustering (KIFCM) and applied it in e-learning customer analysis [4]. Another clustering approach applied in detecting learners' behavioral patterns to support individual and group-based collaborative learning was put forward by Köck and Paramythis [5].

All the above methods combine traditional clustering algorithms and apply them in Web-based learning systems,

where learners' attributes information is extracted consequently through analyzing their learning behaviors and finally utilized for user clustering. Most of these attributes are in the category of continuous data. From learners' behavioral information, we can easily find quite a lot of continuous data, such as "total time length of learning resources" and "comprehensive test result." In contrast, there also exist attributes data with categorical features which is easily neglected, like "chosen lecturer," "chosen learning resource type," and so forth. Although this kind of data is a smaller component of learning behavior information, it also plays a significant role in learner clustering.

In addition, the mixed data of discrete and continuous data, which is extracted from learning behaviors in Web-based learning systems, are interrelated. There are implicit coupling relationships among them. Clustering is often ignored by the traditional clustering algorithms, which leads to massive significant information loss during the process of similarity computation and user clustering. Consequently, the quality of relevant services provided, like learning guides and learning resources recommendation, is not satisfactory. For example, we have the common sense that "total time length of learning resources" has positive impact on "comprehensive test result." Generally, if the "total time length of learning resources" is longer, the "comprehensive test result" is better. However, there are also some special groups of students who behave differently. They can either get better "comprehensive test result" with shorter "total time length of learning resources" or worse "comprehensive test result" with longer "total time length of learning resources." The special correlation between attributes, which is often ignored, is considered in user clustering of our approach. This will lead to certain effect on user clustering accuracy but will not lead to guaranteeing that all users can get highly qualified personalized services easily. An effect mechanism is needed to respond to the loss of the ignored information.

To solve the above issues, this paper proposed a coupled user clustering algorithm based on mixed data, namely, CUCA-MD. This algorithm is based on the truth that both discrete and continuous data exist in learning behavior information; it, respectively, analyzes them according to their different features. In the analysis, CUCA-MD fully takes into account intracoupling and intercoupling relationships and builds user similarity matrixes, respectively, for discrete attribute and continuous attributes. Ultimately we get the integrated similarity matrix using weighted summation and implement user clustering with the help of spectral clustering algorithm. In this way, we take full advantage of the mixed data generated from learning actions in Web-based learning systems. Meanwhile, the algorithm well considers the correlation and coupling relationships of attributes, which enables us to find interactions between users, especially users of previously mentioned special groups. Consequently it can provide suitable and efficient learning guidance and help for users.

The contributions of this algorithm can be summarized from three aspects. Firstly, it takes into account the coupling relationships of attributes in Web-based learning systems, which is frequently neglected before, and improves clustering

accuracy. Secondly, it fully considers different features of discrete data and continuous data and builds user similarity matrix based on mixed data. Thirdly, it captures and analyzes individuals' learning behaviors and provides customized and personalized learning services to different groups of learners.

The rest of the paper is organized as follows. The next section introduces related works. The clustering algorithm model is proposed in Section 3. Section 4 introduces detailed utilization of the clustering algorithm. Discrete and continuous data analysis are also studied in this section. In Section 5, experiments and results analysis are demonstrated. Section 6 concludes this paper.

2. Related Works

Using mixed data to do user clustering has been achieved in some fields but rarely in Web-based learning area. Ahmad and Dey came up with a clustering algorithm based on updated k -mean paradigm which overcomes the numeric data only limitation, which works well for data with mixed numeric and categorical features [6]. A k -prototypes algorithm was proposed, defined as a combined dissimilarity measure, and further integrates the k -means which deals with numeric data and k -modes algorithm which uses a simple matching dissimilarity measure to deal with categorical objects, to allow for clustering objects described by mixed numeric and categorical attributes [7]. Another automated technique, called *SpectralCAT*, was addressed for unsupervised clustering of high-dimensional data that contains numerical or nominal or mix of attributes, suggesting automatically transforming the high-dimensional input data into categorical values [8].

Recently an increasing number of researchers pay special attention to interactions of object attributes and have been aware that the independence assumption on attributes often leads to a mass of information loss. In addition to the basic Pearson's correlation [9], Wang et al. addressed intracoupled and intercoupled interactions of continuous attributes [10], while Li et al. proposed an innovative coupled group-based matrix factorization model for discrete attributes of recommender system [11]. An algorithm to detect interactions between attributes was addressed, but it is only applicable in supervised learning with the experimental results [12]. Calders et al. proposed the use of rank based measures to score the similarity of sets of numerical attributes [13]. Bollegala et al. proposed method comprises two stages: learning a lower-dimensional projection between different relations and learning a relational classifier for the target relation type with instance sampling [14]. From all our viewed documents, we hardly find anything of taking into account coupling relationships of user attributes for user clustering in Web-based learning systems.

3. Clustering Model

User clustering model plays a significant role in user evaluation framework [15]. In this section, the coupled user clustering model based on mixed data is illustrated in

FIGURE 1: The coupled user clustering model based on mixed data.

Figure 1. This model, respectively, takes into account the discrete data and continuous data generated in learning behaviors and incorporates intracoupled and intercoupled relationships of attributes in user clustering. Compared with traditional algorithms, it captures the hidden user interaction information by fully analyzing mixed data, which improves clustering accuracy.

The model is built on the basis of the discrete data and continuous data extracted from learning behaviors. According to their different features, we tease out the corresponding attributes in parallel through analyzing the behaviors. Then intracoupled and intercoupled relationships are introduced into user similarity computation, which helps to get user similarity matrixes, respectively, for discrete attributes and continuous attributes. Finally we use weighted summation to integrate the two matrixes and apply Ng-Jordan-Weiss (NJW) spectral clustering algorithm [16] in user clustering. With the clustering result applied in Web-based learning systems, various personalized services are provided accordingly,

such as learning strategy customization, learning tutoring, and learning resources recommendation.

4. Clustering Algorithm

This paper proposed a coupled user clustering algorithm based on mixed data, which is suitable to be applied in education field. It fits for not only user clustering analysis in Web-based learning systems but also corporate training and performance review, as well as other Web-based activities, in which user participation and behavior recording is involved. The implementation of the CUCA-MD in Web-based learning systems is introduced in this section.

4.1. Discrete Data Analysis. Among the data generated from users' learning behaviors, discrete data plays a significant role in user behavior analysis and user clustering. In the following section, the procedure of how to compute user

TABLE 1: A fragment example of user discrete attributes.

U	A		
	a_1	a_2	a_3
u_1	Wang	Video	Online
u_2	Liu	Video	Online
u_3	Zhao	e-book	Interactive
u_4	Li	Audio	Offline
u_5	Li	PPT	Offline

similarity using discrete data in Web-based learning systems is demonstrated, during which intracoupled similarity within an attribute (i.e., value frequency distribution) and intercoupled similarity between attributes (i.e., feature dependency aggregation) are also considered.

4.1.1. User Learning Behavior Analysis. In Web-based learning systems (<https://www.khanacademy.org/>), (<https://www.coursera.org/>), usually various discrete data will be generated during learning process, such as chosen lecturer, chosen learning resource type and chosen examination form, evaluation on lecturer and learning content, main learning time period, and uploading and downloading learning resources. To make it more explicit, we choose “chosen lecturer,” “chosen learning resource type,” and “chosen examination form” as the attributes for later analysis, respectively, denoted by a_1 , a_2 , and a_3 ; we also choose 5 students as the study objects, denoted by u_1 , u_2 , u_3 , u_4 , and u_5 . The objects and their attributes values are shown in Table 1. Thus, we discuss the similarity of categorical values by considering data characteristics. Two attribute values are similar if they present analogous frequency distributions for one attribute [17]; this reflects the intracoupled similarity within a feature. In Table 1, for example, “Wang,” “Liu,” and “Zhao” are similar because each of them appears once. However, the reality is “Wang” and “Liu” are more similar than “Wang” and “Zhao,” because the “chosen learning resource type” and “chosen examination form” of lecturer “Wang” and “Liu” are identical. If we need to recommend a lecturer to students who like Wang’s lectures more, we will prefer Liu instead of Zhao because Liu’s lecture is more easily accepted by students. It indicates that the similarity between “chosen lecturer” should also cater for the dependencies on other features such as “chosen learning resources” and “chosen examination form” over all objects, namely, the intercoupled similarity between attributes.

4.1.2. Intracoupled and Intercoupled Representation. Data objects with features can be organized by the information table $S = \langle U, A, V, f \rangle$, where $U = \{u_1, u_2, \dots, u_m\}$ is composed of a nonempty finite number of users, $A = \{a_1, a_2, \dots, a_n\}$ is a finite set of discrete attributes, $V = \bigcup_{j=1}^n V_j$ is a set of all attribute values, $V_j = \{a_j \cdot v_1, \dots, a_j \cdot v_{t_j}\}$ is a set of attribute values of the j th attribute, namely, a_j ($1 \leq j \leq n$), and $f = \bigcup_{j=1}^n f_j$, $f_j : U \rightarrow V_j$ is an information function which assigns a particular value of each feature to every user. We take Table 2 as an example to explicitly

TABLE 2: An example of information table.

U	A		
	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3

represent intracoupled and intercoupled similarity of discrete attributes.

To analyze intracoupled and intercoupled correlation of user attributes, we define a few basic concepts as follows.

Definition 1. Given an information table S , 3 set information functions (SIFs) are defined as follows:

$$\begin{aligned} f_j^* (\{u_{k_1}, \dots, u_{k_i}\}) &= \{f_j(u_{k_1}), \dots, f_j(u_{k_i})\}, \\ g_j(x) &= \{u_i \mid f_j(u_i) = x, 1 \leq j \leq n, 1 \leq i \leq m\}, \\ g_j^*(W) &= \{u_i \mid f_j(u_i) \in W, 1 \leq j \leq n, 1 \leq i \leq m\}, \end{aligned} \quad (1)$$

where f^* is the mapping function of user set to attribute values, g is mapping function of attribute values to user, and g^* is mapping function of attribute value set to user. $u_i, u_{k_1}, \dots, u_{k_i} \in U$, $W \subseteq V_j$, and j denotes the j th attribute.

These SIFs describe the relationships between objects and attribute values from different levels. For example, $f_2^* (\{u_1, u_2, u_3\}) = \{B_1, B_2\}$ and $g_2(B_1) = \{u_1, u_2\}$ for value B_1 , while $g_2^* (\{B_1, B_2\}) = \{u_1, u_2, u_3\}$ if given that $W = \{B_1, B_2\}$.

Definition 2. Given an information table S , an Interinformation Function (IIF) is defined as

$$\varphi_{j \rightarrow k}(x) = f_k^*(g_j(x)). \quad (2)$$

This IIF $\varphi_{j \rightarrow k}$ is the composition of f_k^* and g_j . It obtains the k th attribute value subset for the corresponding objects, which are derived from j th attribute value x . For example, $\varphi_{2 \rightarrow 1}(B_1) = \{A_1, A_2\}$.

Definition 3. Given an information table S , the k th attribute value subset $W \subseteq V_k$, and the j th attribute value $x \in V_j$, the Information Conditional Probability (ICP) of W with respect to x is defined as

$$P_{k|j}(W \mid x) = \frac{|g_k^*(W) \cap g_j(x)|}{|g_j(x)|}. \quad (3)$$

When given all the objects with the j th attribute value x , ICP means the percentage of users whose k th attributes fall in subset W and j th attribute value is x as well. For example, $P_{1|2}(\{A_1\} \mid B_1) = 0.5$.

Intracoupled and intercoupled similarity of attributes are, respectively, introduced as follows.

Intracoupled Interaction. Based on [9], intracoupled similarity is decided by attribute value occurrence times in terms of frequency distribution. When we calculate an attribute's intracoupled similarity, we consider the relationship between attribute value frequencies on one feature, demonstrated as follows.

Definition 4. Given an information table S , the Intracoupled Attribute Value Similarity (IaAVS) between attribute values x and y of features a_j is defined as

$$\delta_j^{\text{Ia}}(x, y) = \frac{|g_j(x)| \cdot |g_j(y)|}{|g_j(x)| + |g_j(y)| + |g_j(x)| \cdot |g_j(y)|}. \quad (4)$$

Greater similarity is assigned to the attribute value pair which owns approximately equal frequencies. The higher these frequencies are, the closer such two values are. For example, $\delta_2^{\text{Ia}}(B_1, B_2) = 0.4$.

Intercoupled Interaction. We have considered the intracoupled similarity, that is, the interaction of attribute values within one feature a_j . However, this does not cover interaction between different attributes, namely, a_k ($k \neq j$) and a_j .

Cost and Salzberg [18] came up with a method which is for measuring the overall similarities of classification of all objects on each possible value of each feature. If attributes values occur with the same relative frequency for all classifications, they are identified as being similar. This interaction between features in terms of cooccurrence is taken as intercoupled similarity.

Definition 5. Given an information table S , the intercoupled relative similarity based on Intersection Set (IRSI) between different values x and y of feature a_j regarding another feature a_k is formalized as

$$\delta_{j|k}(x, y) = \sum_{\omega \in \cap} \min \{P_{k|j}(\{\omega\} | x), P_{k|j}(\{\omega\} | y)\}, \quad (5)$$

where $\omega \in \cap$ denote $\omega \in \varphi_{j \rightarrow k}(x) \cap \varphi_{j \rightarrow k}(y)$, respectively.

The value subset $W \subseteq V_k$ is replaced with $\omega \in V_k$, which is considered to simplify computation.

With (5), for example, the calculation of $\delta_{2|1}(B_1, B_2)$ is much simplified since only $A_2 \in \varphi_{2 \rightarrow 1}(B_1) \cap \varphi_{2 \rightarrow 1}(B_2)$; then we can easily get $\delta_{2|1}(B_1, B_2) = 0.5$. Thus, this method is quite efficient in reducing intracoupled relative similarity complexity.

Definition 6. Given an information table S , the Intercoupled Attribute Value Similarity (IeAVS) between attribute values x and y of feature a_j is defined as

$$\delta_j^{\text{Ie}}(x, y) = \sum_{k=1, k \neq j}^n \alpha_k \delta_{j|k}(x, y), \quad (6)$$

where α_k is the weight parameter for feature a_k , $\sum_{k=1}^n \alpha_k = 1$, $\alpha_k \in [0, 1]$, and $\delta_{j|k}(x, y)$ is one of the intercoupled relative similarity candidates.

In Table 2, for example, $\delta_2^{\text{Ie}}(B_1, B_2) = 0.5 \cdot \delta_{2|1}(B_1, B_2) + 0.5 \cdot \delta_{2|3}(B_1, B_2) = (0.5 + 0)/2 = 0.25$ if $\alpha_1 = \alpha_3 = 0.5$ is taken with equal weight.

4.1.3. Integrated Coupling Representation. Coupled Attribute Value Similarity (CAVS) is proposed in terms of both intracoupled and intercoupled value similarities. For example, the coupled interaction between B_1 and B_2 covers both the intracoupled relationship specified by the occurrence times of values B_1 and B_2 , 2 and 2, and the intercoupled interaction triggered by the other two features, a_1 and a_3 .

Definition 7. Given an information table S , the Coupled Attribute Value Similarity (CAVS) between attribute values x and y of feature a_j is defined as

$$\delta_j^{\text{A}}(x, y) = \delta_j^{\text{Ia}}(x, y) \cdot \delta_j^{\text{Ie}}(x, y), \quad (7)$$

where δ_j^{Ia} and δ_j^{Ie} are IaAVS and IeAVS, respectively.

In Table 2, for instance, CAVS is obtained as $\delta_2^{\text{A}}(B_1, B_2) = \delta_2^{\text{Ia}}(B_1, B_2) \cdot \delta_2^{\text{Ie}}(B_1, B_2) = 0.4 \times 0.25 = 0.1$.

With the specification of IaAVS and IeAVS, a coupled similarity between objects is built based on CAVS. Then we sum all CAVSs analogous to the construction of Manhattan dissimilarity [9]. Formally, we have the following definition.

Definition 8. Given an information table S , the Coupled User Similarity (CUS) between users u_{i_1} and u_{i_2} is defined as

$$\text{CUS}(u_{i_1}, u_{i_2}) = \sum_{j=1}^n \beta_j \delta_j^{\text{A}}(x_{i_1j}, x_{i_2j}), \quad (8)$$

where β_j is the weight parameter of attribute a_j , $\sum_{j=1}^n \beta_j = 1$, $\beta_j \in [0, 1]$, x_{i_1j} and x_{i_2j} are the attribute values of feature a_j for u_{i_1} and u_{i_2} , respectively, and $1 \leq i_1, i_2 \leq m$ and $1 \leq j \leq n$.

In Table 2, for example, $\text{CUS}(u_2, u_3) = \sum_{j=1}^3 \beta_j \delta_j^{\text{A}}(x_{2j}, x_{3j}) = 1/3 \cdot \delta_1^{\text{A}}(A_2, A_2) + 1/3 \cdot \delta_2^{\text{A}}(B_1, B_2) + 1/3 \cdot \delta_3^{\text{A}}(C_1, C_2) = (0.5 + 0.1 + 0.125)/3 \approx 0.24$ if $\beta_1 = \beta_2 = \beta_3 = 1/3$ is taken with equal weight.

In this way, a user similarity matrix of $m \times m$ entries regarding discrete data can be built as

$$M^{\text{DD}} = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \cdots & \cdots & \ddots & \cdots \\ S_{m1} & S_{m2} & \cdots & S_{mm} \end{pmatrix}, \quad (9)$$

where $S_{ii} = 1$, $1 \leq i \leq m$.

For instance, we get a user similarity matrix of 5×5 entries regarding discrete data based on Table 2, as

$$\begin{pmatrix} 1 & 0.40 & 0.14 & 0.04 & 0 \\ 0.40 & 1 & 0.24 & 0.08 & 0 \\ 0.14 & 0.24 & 1 & 0.23 & 0.07 \\ 0.04 & 0.08 & 0.23 & 1 & 0.26 \\ 0 & 0 & 0.07 & 0.26 & 1 \end{pmatrix}. \quad (10)$$

4.2. Continuous Data Analysis. Continuous data is with different features when compared with discrete data. In the following section, user similarity computation is demonstrated using Taylor-like expansion, with the involvement of intracoupled interaction within an attribute (i.e., the correlations between attributes and their own powers) and intercoupled interaction among different attributes (i.e., the correlations between attributes and the powers of others).

4.2.1. User Learning Behavior Analysis. After students log onto a Web-based learning system, the system will record their activity information, such as times of doing homework and number of learning resources. This paper refers to a Web-based personalized user evaluation model [19] and utilizes its evaluation index system to extract students' continuous attributes information. This index system, as shown in Table 3, is based on evaluation standards of America $K-12$ (kindergarten through twelfth grade) [20] and Delphi method [21], which is a hierarchical structure built according to mass of information and data generated during general e-learning activities. It is defined with 20 indicators and can comprehensively represent the students' attributes. Due to the different units used for measuring extracted attributes, like times, time length, amount, percentage, and so forth, we need to normalize them firstly: result is shown in Table 4.

4.2.2. Intracoupled and Intercoupled Representation. In this section, intracoupled and intercoupled relationships of above extracted continuous attributes are, respectively, represented. Here we use an example to make it more explicate. We single out 6 attributes data with continuous feature of the same 5 students mentioned in Section 4.1.1, including "average correct rate of homework," "times of doing homework," "number of learning resources," "total time length of learning resources," "daily average quiz result," and "comprehensive test result," denoted by $a'_1, a'_2, a'_3, a'_4, a'_5$, and a'_6 in the following representations, shown in Table 4.

Here we use an information table $S' = \langle U, A', V', f' \rangle$ to represent user attributes information. $U = \{u_1, u_2, \dots, u_m\}$ means a finite set of users; $A' = \{a'_1, a'_2, \dots, a'_n\}$ refers to a finite set of continuous attributes; $V' = \bigcup_{j=1}^n V'_j$ represents all attributes value sets; $V'_j = \{a'_j \cdot v'_1, \dots, a'_j \cdot v'_{t_j}\}$ is the value set of the j th attribute; $f' = \bigcup_{j=1}^n f'_j$, $f'_j : U \rightarrow V'_j$ is the function for calculating a certain attribute value. For example, the information in Table 4 contains 5 users $\{u_1, u_2, u_3, u_4, u_5\}$ and 6 attributes $\{a'_1, a'_2, a'_3, a'_4, a'_5, a'_6\}$; the first attribute value of u_1 is 0.61.

TABLE 3: Comprehensive evaluation index system.

First-level index	Second-level index
Autonomic learning	Times of doing homework
	Average correct rate of homework
	Number of learning resources
	Total time length of learning resources
	Times of daily quiz
	Daily average quiz result
	Comprehensive test result
	Number of collected resources
	Times of downloaded resources
	Times of making notes
Interactive learning	Times of asking questions
	Times of marking and remarking
	Times of answering classmates' questions
	Times of posting comments on the BBS
	Times of interaction by BBS message
	Times of sharing resources
	Average marks made by the teacher
	Average marks made by other students
	Times of marking and remarking made by the student for the teacher
	Times of marking and remarking made by the student for other students

TABLE 4: A fragment example of user continuous attributes.

U	A'					
	a'_1	a'_2	a'_3	a'_4	a'_5	a'_6
u_1	0.61	0.55	0.47	0.72	0.63	0.62
u_2	0.75	0.92	0.62	0.63	0.74	0.74
u_3	0.88	0.66	0.71	0.74	0.85	0.87
u_4	0.24	0.83	0.44	0.29	0.21	0.22
u_5	0.93	0.70	0.66	0.81	0.95	0.93

The usual way to calculate the interactions between 2 attributes is Pearson's correlation coefficient [9]. For instance, the Pearson's correlation coefficient between a'_k and a'_j is formalized as

$$\begin{aligned} \text{Cor}(a'_j, a'_k) &= \frac{\sum_{u \in U} (f'_j(u) - \mu_j)(f'_k(u) - \mu_k)}{\sqrt{\sum_{u \in U} (f'_j(u) - \mu_j)^2 \sum_{u \in U} (f'_k(u) - \mu_k)^2}}, \end{aligned} \quad (11)$$

where μ_j and μ_k are, respectively, mean values of a'_j and a'_k .

However, the Pearson's correlation coefficient only fits for linear relationship. It is far from sufficient to fully capture pairwise attributes interactions. Therefore we expect to use more dimensions to expand the numerical space spanned by n' and then expose attributes coupling relationship by exploring updated attributes interactions [22].

TABLE 5: Extended user continuous attributes.

	$\langle a'_1 \rangle^1$	$\langle a'_1 \rangle^2$	$\langle a'_2 \rangle^1$	$\langle a'_2 \rangle^2$	$\langle a'_3 \rangle^1$	$\langle a'_3 \rangle^2$	$\langle a'_4 \rangle^1$	$\langle a'_4 \rangle^2$	$\langle a'_5 \rangle^1$	$\langle a'_5 \rangle^2$	$\langle a'_6 \rangle^1$	$\langle a'_6 \rangle^2$
u_1	0.61	0.37	0.55	0.30	0.47	0.22	0.72	0.52	0.63	0.40	0.62	0.38
u_2	0.75	0.56	0.92	0.85	0.62	0.38	0.63	0.40	0.74	0.55	0.74	0.55
u_3	0.88	0.77	0.66	0.44	0.71	0.50	0.74	0.56	0.85	0.72	0.87	0.76
u_4	0.24	0.06	0.83	0.69	0.44	0.19	0.29	0.08	0.21	0.04	0.22	0.05
u_5	0.93	0.86	0.70	0.49	0.66	0.44	0.81	0.66	0.95	0.90	0.93	0.86

Firstly, we use a few additional attributes to expand interaction space in the original information table. Hence, there are L attributes for each original attribute a'_j , including itself, namely, $\langle a'_j \rangle^1, \langle a'_j \rangle^2, \dots, \langle a'_j \rangle^L$. Each attribute value is the power of the attribute; for instance, $\langle a'_j \rangle^3$ is the third power of attribute a'_j and $\langle a'_j \rangle^p$ ($1 \leq p \leq L$) is the p th power of a'_j . In Table 4, the denotation a'_j and $\langle a'_j \rangle^1$ are equivalent; the value of $\langle a'_j \rangle^2$ is the square of a'_j value. For simplicity, we set $L = 2$ in Table 5.

Secondly, the correlation between pairwise attributes is calculated. It captures both local and global coupling relations. We take the p values for testing the hypotheses of no correlation between attributes into account. p value here means the probability of getting a correlation as large as possible observed by random chance, while the true correlation is zero. If p value is smaller than 0.05, the correlation $\text{Cor}(a'_j, a'_k)$ is significant. The updated correlation coefficient is as follows:

$$R\text{-Cor}(a'_j, a'_k) = \begin{cases} \text{Cor}(a'_j, a'_k) & \text{if } p\text{-value} < 0.05, \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

Here we do not consider all relationships but only take the significant coupling relationships into account, because all relationships involvement may cause the overfitting issue on modeling coupling relationship. This issue will go against the attribute inherent interaction mechanism. So based on the updated correlation, the intracoupled and intercoupled interaction of attributes are proposed. Intracoupled interaction is the relationship between a'_j and all its powers; intercoupled interaction is the relationship between a'_j and powers of the rest of the attributes a'_k ($k \neq j$).

Intracoupled Interaction. The intracoupled interaction within an attribute is represented as a matrix. For attribute a'_j , it is $L \times L$ matrix $R^{\text{Ia}}(a'_j)$. In the matrix, (p, q) is the correlation between $\langle a'_j \rangle^p$ and $\langle a'_j \rangle^q$ ($1 \leq p, q \leq L$). Consider

$$R^{\text{Ia}}(a'_j) = \begin{pmatrix} \alpha_{11}(j) & \alpha_{12}(j) & \cdots & \alpha_{1L}(j) \\ \alpha_{21}(j) & \alpha_{22}(j) & \cdots & \alpha_{2L}(j) \\ \cdots & \cdots & \ddots & \cdots \\ \alpha_{L1}(j) & \alpha_{L2}(j) & \cdots & \alpha_{LL}(j) \end{pmatrix}, \quad (13)$$

where $\alpha_{pq}(j) = R\text{-Cor}(\langle a'_j \rangle^p, \langle a'_j \rangle^q)$ is the Pearson's correlation coefficient between $\langle a'_j \rangle^p$ and $\langle a'_j \rangle^q$.

For attribute a'_1 in Table 5, we can get the intracoupled interaction within it as $R^{\text{Ia}}(a'_1) = \begin{pmatrix} 1 & 0.986 \\ 0.986 & 1 \end{pmatrix}$, which means that the correlation coefficient between attribute "average correct rate of homework" and its second power is as high as 0.986. There is close relationship between them.

Intercoupled Interaction. The intercoupled interaction between attribute a'_j and other attributes a'_k ($k \neq j$) is quantified as $L \times L * (n' - 1)$ matrix as

$$\begin{aligned} & R^{\text{Ie}}(a'_j | \{a'_k\}_{k \neq j}) \\ &= (R^{\text{Ie}}(a'_j | a'_{k_1}) \ R^{\text{Ie}}(a'_j | a'_{k_2}) \ \cdots \ R^{\text{Ie}}(a'_j | a'_{k_{n'-1}})), \\ & R^{\text{Ie}}(a'_j | a'_{k_i}) \\ &= \begin{pmatrix} \beta_{11}(j | k_i) & \beta_{12}(j | k_i) & \cdots & \beta_{1L}(j | k_i) \\ \beta_{21}(j | k_i) & \beta_{22}(j | k_i) & \cdots & \beta_{2L}(j | k_i) \\ \cdots & \cdots & \ddots & \cdots \\ \beta_{L1}(j | k_i) & \beta_{L2}(j | k_i) & \cdots & \beta_{LL}(j | k_i) \end{pmatrix}. \end{aligned} \quad (14)$$

Here $\{a'_k\}_{k \neq j}$ refers to all the attributes except for a'_j , and $\beta_{pq}(j | k_i) = R\text{-Cor}(\langle a'_j \rangle^p, \langle a'_{k_i} \rangle^q)$ is the correlation coefficient between $\langle a'_j \rangle^p$ and $\langle a'_{k_i} \rangle^q$ ($1 \leq p, q \leq L$).

For attribute a'_1 in Table 5, the intercoupled interaction between a'_1 and others ($a'_2, a'_3, a'_4, a'_5, a'_6$) is calculated as

$$\begin{aligned} & R^{\text{Ie}}(a'_1 | \{a'_2, a'_3, a'_4, a'_5, a'_6\}) \\ &= \begin{pmatrix} 0 & 0 & 0.898 & 0.885 & 0.928 & 0.921 & 0.997 & 0.982 & 0.999 & 0.988 \\ 0 & 0 & 0.929 & 0.920 & 0.879 & 0.888 & 0.978 & 0.994 & 0.982 & 0.999 \end{pmatrix}. \end{aligned} \quad (15)$$

The p values between a'_1 and others ($a'_2, a'_3, a'_4, a'_5, a'_6$) are calculated as

$$p^{le}(a'_1 | \{a'_2, a'_3, a'_4, a'_5, a'_6\}) = \begin{pmatrix} 0.689 & 0.677 & 0.039 & 0.046 & 0.023 & 0.027 & 0 & 0.003 & 0 & 0.002 \\ 0.733 & 0.707 & 0.023 & 0.027 & 0.050 & 0.044 & 0.004 & 0.001 & 0.003 & 0 \end{pmatrix}. \quad (16)$$

Based on the result, we can find that there is hidden correlation between user attributes. For instance, all the p values between attributes a'_1 and a'_2 are larger than 0.05, so the correlation coefficient is 0 based on (12), indicating there is no significant correlation between “average correct rate of homework” and “times of doing homework.” Meanwhile, the correlation coefficient between a'_1 and a'_5 and a'_1 and a'_6 is quite close to 1; it indicates that “daily average quiz result” and “comprehensive test result” have close relationship, respectively, with “average correct rate of homework,” which is consistent with our practical experiences. In conclusion, comprehensively taking into account intracoupled and intercoupled correlation of attributes can efficiently help capturing relationships between user attributes.

4.2.3. Integrated Coupling Representation. Intracoupled and intercoupled interactions are integrated in this section as a coupled representation scheme.

In Table 5, each user is signified by $L * n'$ updated variables $\widetilde{A}' = \{\langle a'_1 \rangle^1, \dots, \langle a'_1 \rangle^L, \dots, \langle a'_{n'} \rangle^1, \dots, \langle a'_{n'} \rangle^L\}$. With the updated function $\widetilde{f}_j^p(u)$, the corresponding value of attribute $\langle a'_j \rangle^p$ is assigned to user u . Attribute a'_j and all its powers are signified as $\widetilde{u}(a'_j) = [\widetilde{f}_j^1(u), \dots, \widetilde{f}_j^L(u)]$, while the rest of the attributes and all powers are presented in another vector $\widetilde{u}(\{a'_k\}_{k \neq j}) = [\widetilde{f}_{k_1}^1(u), \dots, \widetilde{f}_{k_1}^L(u), \dots, \widetilde{f}_{k_{n'-1}}^1(u), \dots, \widetilde{f}_{k_{n'-1}}^L(u)]$. For instance, in Table 5, $\widetilde{u}_1(a'_1) = [0.61, 0.37]$, $\widetilde{u}_1(\{a'_2, a'_3, a'_4, a'_5, a'_6\}) = [0.55, 0.30, 0.47, 0.22, 0.72, 0.52, 0.63, 0.40, 0.62, 0.38]$.

Definition 9. The coupled representation of attribute a'_j is formalized as $1 \times L$ vector $u^c(a'_j | \widetilde{A}', L)$, where $(1, p)$ component corresponds to the updated attribute $\langle a'_j \rangle^p$. One has

$$u^c(a'_j | \widetilde{A}', L) = u^{la}(a'_j | \widetilde{A}', L) + u^{le}(a'_j | \widetilde{A}', L), \quad (17)$$

$$u^{la}(a'_j | \widetilde{A}', L) = \widetilde{u}(a'_j) \odot w \otimes [R^{la}(a'_j)]^T, \quad (18)$$

$$u^{le}(a'_j | \widetilde{A}', L) = \widetilde{u}(\{a'_k\}_{k \neq j}) \odot [w, w, \dots, w] \otimes [R^{le}(a'_j | \{a'_k\}_{k \neq j})]^T, \quad (19)$$

where $w = [1/(1!), 1/(2!), \dots, 1/(L!)]$ is a constant $1 \times L$ vector, $[w, w, \dots, w]$ is a $1 \times L * (n' - 1)$ vector concatenated by $n' - 1$ constant vectors w . $\odot$ denotes the Hadamard product, and $\otimes$ represents the matrix multiplication.

Taking an example in Table 6, the coupled representation for attribute a'_1 is presented as $u^c(a'_1 | \widetilde{A}', 2) = [3.85, 3.80]$. The reason why we choose such a representation method is explained below. If (17) is expanded, for example, we get the $(1, p)$ element which corresponds to $\langle a'_j \rangle^p$ of the vector $u^c(a'_j | \widetilde{A}', L)$ as below, which resembles Taylor-like expansion of functions [23]

$$\begin{aligned} u^c(a'_j | \widetilde{A}', L) \cdot \langle a'_j \rangle^p &= \alpha_{p1}(j) \cdot \widetilde{f}_j^1(u) + \sum_{i=1}^{n'-1} \frac{\beta_{p1}(j | k_i)}{1!} \widetilde{f}_{k_i}^1(u) \\ &+ \frac{\alpha_{p2}(j)}{2!} \widetilde{f}_j^2(u) + \sum_{i=1}^{n'-1} \frac{\beta_{p2}(j | k_i)}{2!} \widetilde{f}_{k_i}^2(u) + \dots \\ &+ \frac{\alpha_{pL}(j)}{L!} \widetilde{f}_j^L(u) + \sum_{i=1}^{n'-1} \frac{\beta_{pL}(j | k_i)}{L!} \widetilde{f}_{k_i}^L(u). \end{aligned} \quad (20)$$

Finally we obtained the global coupled representation of all the n' original attributes as a concatenated vector:

$$\begin{aligned} u^c(\widetilde{A}', L) &= [u^c(a'_1 | \widetilde{A}', L), u^c(a'_2 | \widetilde{A}', L), \dots, \\ &u^c(a'_{n'} | \widetilde{A}', L)]. \end{aligned} \quad (21)$$

Incorporated with the couplings of attributes, each user is represented as $1 \times L * n'$ vector. When all the users follow the steps above, we then obtain $m \times L * n'$ coupled information table. For example, based on Table 4, the coupled information table shown in Table 6 is the new representation.

With the new user attributes information of the coupled information table, we utilize the formula below [16] to compute user similarity and build a matrix M^{CD} of $m \times m$ entries

$$S'_{ij} = \exp\left(-\frac{\|u_i - u_j\|^2}{2\sigma^2}\right), \quad (22)$$

where $i \neq j$, $S'_{ii} = 1$, and σ denotes scaling parameter. Here we take $\sigma = 0.3$. Detailed parameter estimation procedure is introduced in experiment 5.2

$$M^{CD} = \begin{pmatrix} S'_{11} & S'_{12} & \dots & S'_{1m} \\ S'_{21} & S'_{22} & \dots & S'_{2m} \\ \dots & \dots & \ddots & \dots \\ S'_{m1} & S'_{m2} & \dots & S'_{mm} \end{pmatrix}. \quad (23)$$

TABLE 6: Integrated coupling representation of user continuous attributes.

	$\langle a'_1 \rangle^1$	$\langle a'_1 \rangle^2$	$\langle a'_2 \rangle^1$	$\langle a'_2 \rangle^2$	$\langle a'_3 \rangle^1$	$\langle a'_3 \rangle^2$	$\langle a'_4 \rangle^1$	$\langle a'_4 \rangle^2$	$\langle a'_5 \rangle^1$	$\langle a'_5 \rangle^2$	$\langle a'_6 \rangle^1$	$\langle a'_6 \rangle^2$
u_1	3.85	3.80	0.70	0.70	2.20	1.46	3.24	3.23	3.35	3.70	3.76	3.81
u_2	4.54	4.50	1.34	1.34	2.89	1.98	3.66	3.65	3.82	4.31	4.37	4.51
u_3	5.51	5.46	0.88	0.88	3.54	2.44	4.46	4.45	4.66	5.22	5.28	5.47
u_4	1.53	1.52	1.17	1.17	1.01	0.80	1.03	1.02	1.06	1.42	1.44	1.52
u_5	5.94	5.89	0.94	0.94	3.73	2.49	4.95	4.94	5.17	5.68	5.75	5.90

For instance, we get a user similarity matrix of 5×5 entries regarding continuous data based on Table 4, as

$$\begin{pmatrix} 1 & 0.11 & 0.20 & 0.08 & 0.14 \\ 0.11 & 1 & 0.20 & 0.06 & 0.21 \\ 0.20 & 0.20 & 1 & 0.03 & 0.70 \\ 0.08 & 0.06 & 0.03 & 1 & 0.02 \\ 0.14 & 0.21 & 0.70 & 0.02 & 1 \end{pmatrix}. \quad (24)$$

4.3. User Clustering. In Sections 4.1 and 4.2, we get separate user similarity matrix M^{DD} regarding discrete attributes and M^{CD} regarding continuous attributes. With weighted summation, an integrated matrix M^{MD} of $m \times m$ entries based on mixed data can be obtained as

$$M^{\text{MD}} = \gamma_1 M^{\text{DD}} + \gamma_2 M^{\text{CD}}, \quad (25)$$

where γ_1 and γ_2 are the respective weights of discrete attributes and continuous attributes. $\gamma_1 + \gamma_2 = 1$, $\gamma_1/\gamma_2 = n/n'$, and n denotes the number of former attributes while n' denotes that of the latter ones.

For example, in the former examples we listed 3 discrete attributes and 6 continuous attributes; then $\gamma_1 = 1/3$ and $\gamma_2 = 2/3$. For users of u_1, u_2, u_3, u_4 , and u_5 , the user similarity matrix based on the mixed data is obtained as follows:

$$\begin{pmatrix} 1 & 0.21 & 0.18 & 0.06 & 0.09 \\ 0.21 & 1 & 0.21 & 0.06 & 0.14 \\ 0.18 & 0.21 & 1 & 0.09 & 0.49 \\ 0.06 & 0.06 & 0.09 & 1 & 0.10 \\ 0.09 & 0.14 & 0.49 & 0.10 & 1 \end{pmatrix}. \quad (26)$$

With consideration of intracoupled and intercoupled correlation of user attributes, we get the user similarity matrix M^{MD} based on mixed learning behavior data. Next, with NJW spectral clustering, user clustering procedure is described. Detailed clustering result is demonstrated in experiments part.

5. Experiments and Evaluation

We conducted experiments and user studies using the coupled user clustering algorithm proposed in this paper. The data for the experiments are collected from a Web-based learning system of China Educational Television (CETV), named “New Media Learning Resource Platform for National

Education” (<http://www.guoshi.com/>). As a basic platform for national lifelong education, which started the earliest in China, has the largest group of users and provides most extensive learning resources, it meets the needs of personalization and diversity of different users through integrating a variety of fusion network, terminals, and resources. So far, the number of registered users has reached more than two million. Experiments are carried out to verify the algorithm’s validity and accuracy. The experiment is composed of four parts: user study, parameter estimation, user clustering, and result analysis.

5.1. User Study. In the experiment, we asked 180 users (indicated by $s_1, s_2, \dots, s_{220}$) to learn Data Structures online. The whole learning process, including the recording and analysis of learning activities information, was accomplished in CETV mentioned above.

Recently public data sets regarding learners’ learning behaviors in online learning systems are insufficient, and most of them do not contain labeled user clustering information. Meanwhile, because learners’ behaviors are always with certain subjectivity, the accuracy of labeling learners with different classifiers only based on behaviors but without knowing the information behind is not full. Therefore, we adopt a few user studies, directly and, respectively, collecting relevant user similarity data from students and teachers, as the basis for verifying the accuracy of learners clustering in Web-based learning systems.

Through analyzing the continuous attributes extracted from Table 3 according to user evaluation index system, we can easily find that they can be mainly classified to two kinds: one kind of attributes reflecting learners’ learning attitude, like “times of doing homework,” “number of learning resources,” and “total time length of learning resources”; the other kind of attributes reflecting learners’ learning effect, like “average correct rate of homework,” “daily average quiz result,” and “comprehensive test result.” Meanwhile, we also analyze attributes with categorical features, for example, “chosen lecturer,” “chosen learning resource type,” and “chosen examination form,” which all reflect learners’ learning preferences. Therefore, we ask the students and teachers together to comprehensively estimate students’ similarity, respectively, from three perspectives, which are learning attitude, learning effect, and learning preference. We request each of the 180 students to choose the top 5 of other students who are mostly like himself and 5 who are hardly like himself, taking into account the three perspectives; each of the lecturers who are giving the lesson of data structure also makes options for

TABLE 7: The clustering results of s_1 .

		Top 5 “most similar”	Top 5 “least similar”
Before clustering	s_1	$s_7, s_{16}, s_{35}, s_{84}, s_{103}$	$s_{19}, s_{55}, s_{82}, s_{122}, s_{131}$
	Mr. Liu	$s_7, s_{35}, s_{60}, s_{103}, s_{162}$	$s_4, s_{19}, s_{55}, s_{82}, s_{131}$
After clustering	s_1	$s_7, s_{16}, s_{35}, s_{103}$ stay in the same cluster with s_1	None of them stays in the same cluster with s_1
		Similarity accuracy: 80%	Dissimilarity accuracy: 100%
	Mr. Liu	$s_7, s_{35}, s_{103}, s_{162}$ stay in the same cluster with s_1	s_4 stays in the same cluster with s_1
	Comprehensive accuracy	Similarity accuracy: 80%	Dissimilarity accuracy: 80%
		Comprehensive similarity accuracy: 80%	Comprehensive dissimilarity accuracy: 91%

the same request for every student in his class. For instance, student s_1 chooses the lesson of lecturer *Liu* and the options, respectively, made by them are shown in Table 7.

5.2. Parameter Estimation. As indicated in (20), the proposed coupled representation for numerical objects is strongly dependent on the maximal power L . Here, we conduct several experiments to study the performance of L with regard to the clustering accuracy of CUCA-MD. The maximal power L is set to range from $L = 1$ to $L = 10$ since $L!$ becomes extremely large when L grows, which means $L = 10$ is probably large enough to obtain most of the information in (20). The experiment verifies that, with the increasing value of L , the clustering accuracy goes higher. When $L = 3$, it reaches a stable point for accuracy change; when $L = 10$, compared with the former, there is only very tiny improvement of accuracy. Therefore, with the precondition for experiment accuracy, we take $L = 3$, reducing the algorithm complexity as much as possible.

We keep adjusting the number of clusters k with a large number of experiments. Finally we take the number as 6, considering the user features in online learning systems. Besides, (22) is needed when computing user similarity using continuous data. The scaling parameter σ of the equation should be set manually, so we test different σ values to get different clustering accuracy and then pick up the optimal one. In Figure 2, the relation of σ values and clustering accuracy is illustrated. When $\sigma = 0.3$, the clustering accuracy is the best; when $\sigma \geq 0.7$, the clustering results stay in a comparatively stable range with no much difference in between. Thus, we take $\sigma = 0.3$.

5.3. User Clustering. In the following experiments, we take use of the 20 continuous user attributes in Table 3 and the 8 discrete attributes (chosen lecturer, chosen learning resource type, chosen examination form, learning time, evaluation of lecturer, evaluation of learning resources, upload, and download) to do user clustering. Because the procedure of recording and analyzing users' learning behaviors is persistent, we divide the learning process to six phases, namely, 5 h, 10 h, 15 h, 20 h, 25 h, and 30 h. Then with the data of different phase, we do user clustering with different algorithm.

FIGURE 2: Parameter estimation of σ .

To verify the efficiency of CUCA-MD in user clustering in Web-based learning systems, we compare it with three other algorithms, which are also based on mixed data, namely, *k-prototype* [7], *mADD* [6], and *spectralCAT* [8]. Besides, to demonstrate the significance of learning behavior with both categorical feature and continuous feature, as well as the different effect of clustering with and without coupling, we take six different methods to, respectively, do clustering. The first one is Simple Matching Similarity (SMS), which only uses 0s and 1s to distinguish similarities between distinct and identical categorical values [9], used to analyze users' discrete attributes and compute user similarity and then applied in user clustering with the help of NJW. This method is named NJW-DD. The second one is described in Section 4.1, which analyzes users' discrete attributes considering intracoupled and intercoupled relationships and then computes user similarity and does user clustering combined with NJW algorithm. This method is called CUCA-DD. The third one is to get clustering result through analyzing continuous attributes and utilizing NJW, namely, NJW-CD. However, the fourth one takes advantage of users' continuous attributes and their intracoupled and intercoupled correlation to get the user similarity and then, with the help of NJW, get the user clustering result. It is introduced in Section 4.2 already, named CUCA-CD. The fifth method is utilizing

FIGURE 3: Clustering result analysis (30 h).

NJW to do user clustering based on both discrete and continuous attributes but without considering intracoupled and intercoupled correlation of attributes. It is named NJW-MD. The sixth is the one proposed in this paper, CUCA-MD.

With the clustering result, we use statistics to make an analysis. In Table 7, taking student s_1 , for example, 4 of the “top most similar” students chosen by s_1 stay in the same cluster with s_1 himself which indicates the clustering accuracy is 80%; thus, we take it as “similarity accuracy.” In contrast, none of the “top least similar” students chosen by s_1 stays in the same cluster with him, indicating the clustering accuracy is 100%; thus, we take it as “dissimilarity accuracy.” In the same way, we also get two accuracy values based on the options made by s_1 ’s lecturer *Mr. Liu*, 80% and 80%. Taking the weight of s_1 ’s option as 55% while taking *Mr. Liu*’s as 45%, we get the comprehensive accuracy values of 80% and 91%, using weighted summation. In this way, we get a pair of accuracy values for each of the 180 students, verifying the efficiency of the different clustering methods.

5.4. Result Analysis. We do comparison analysis on the clustering results using user similarity accuracy and user dissimilarity accuracy. Figure 3 illustrates the clustering accuracy comparison of CUCA-MD and other algorithms, when the average learning length of the 180 students is 30 h. From Figure 3(a), we can easily find that the accuracy of CUCA-MD is higher than that of others regarding mixed data, no matter on similarity or dissimilarity. In Figure 3(b), the clustering results of NJW-DD, CUCA-DD, NJW-CD, CUCA-CD, NJW-MD, and CUCA-MD are demonstrated and compared. We observe that both of the similarity accuracy and dissimilarity accuracy of NJW-DD are the lowest, respectively, 26.2% and 32.3%, when compared with others, while those of CUCA-MD are the highest. Meanwhile, algorithms considering coupling relationship have higher accuracy than

NJW which does not consider it, regardless of discrete data, continuous data, or mixed data. The comprehensive results above verify the outperformance of algorithms which considers coupling relations of attributes; they can efficiently capture the hidden information of behavior data and greatly improve clustering accuracy.

The collection and analysis of learning behaviors are a persistent action, so we illustrate the relationship between average learning length and user clustering accuracy. From Figures 4(a) and 4(b), we can see that, with the extension of learning time, the clustering accuracies of the algorithms based on mixed data become higher, among which CUCA-MD grows the fastest, especially after 20 h. Figures 4(c) and 4(d) show that all accuracies of the algorithms grow, except for NJW-DD. At the same time, algorithm CUCA which considers coupling relationships grows faster than NJW which does not. With the result above, we make the conclusion that is based on only a few attributes and little extracted information: the clustering accuracy of NJW-DD regarding discrete data is not improved much, even with more time and behavior data, while the clustering accuracy of CUCA-MD regarding mixed data, which considers coupling relationship of attributes, is distinctly improved with the increase of behavior data.

Besides, we can verify clustering accuracy through analyzing the structure of user clustering results. The best performance of a clustering algorithm is reaching the smallest distance within a cluster but the biggest distance between clusters; thus, we utilize the evaluation criteria of Relative Distance (the ratio of average intercluster distance upon average intracluster distance) and Sum Distance (the sum of object distances within all the clusters) to present the distance. The larger Relative Distance is and the smaller Sum Distance is, the better clustering results are. From Figure 5, we can see that the Relative Distance of CUCA-MD is larger

FIGURE 4: Clustering result of different phases.

than that of the other algorithms, while the Sum Distance of CUCA-MD is smaller. It indicates that CUCA-MD regarding mixed data, which also considers coupling relationships, outperforms the rest in terms of clustering structure.

6. Conclusion

We proposed a coupled user clustering algorithm based on Mixed Data for Web-based Learning Systems (CUCA-MD) in this paper, which incorporates intracoupled and intercoupled correlation of user attributes with different features. This algorithm is based on the truth that both discrete and continuous data exist in learning behavior information; it, respectively, analyzes them according to different features. In the analysis, CUCA-MD fully takes into account intracoupling and intercoupling relationships and builds user similarity matrixes, respectively, for discrete attribute and continuous attributes. Ultimately we get the integrated similarity matrix using weighted summation and implement user clustering with the help of spectral clustering algorithm. In experiment part, we verify the outperformance of proposed CUCA-MD

in terms of user clustering in Web-based learning systems, through user study, parameter estimation, user clustering, and result analysis.

In this paper, we analyze discrete data and continuous data generated in online learning systems with different methods and build user similarity matrixes regarding attributes with discrete and continuous features, respectively, which makes the algorithm more complicated. In the following studies, we hope to realize the simultaneous processing continuous data and discrete data, while taking into account coupling correlation of user attributes, which will definitely further improve algorithm efficiency.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Project no. 61370137), the National

FIGURE 5: Clustering structure analysis (30 h).

973 Project of China (no. 2012CB720702), and Major Science and Technology Project of Press and Publication (no. GAPP_ZDKJ_BQ/01).

References

- [1] S. Cai and W. Zhu, "The impact of an online learning community project on university Chinese as a foreign language students motivation," *Foreign Language Annals*, vol. 45, no. 3, pp. 307–329, 2012.
- [2] M. Ghosh, "MOOC M4D: an overview and learner's viewpoint on autumn 2013 course," *International Journal of Interactive Mobile Technologies*, vol. 8, no. 1, pp. 46–50, 2014.
- [3] F. Trif, C. Lemnaru, and R. Potolea, "Identifying the user typology for adaptive e-learning systems," in *Proceedings of the 17th IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR '10)*, vol. 3, pp. 1–6, IEEE, May 2010.
- [4] K.-P. Lin, C.-L. Lin, K.-C. Hung, Y.-M. Lu, and P.-F. Pai, "Developing kernel intuitionistic fuzzy c-means clustering for e-learning customer analysis," in *Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM '12)*, pp. 1603–1607, IEEE, December 2012.
- [5] M. Köck and A. Paramythis, "Towards adaptive learning support on the basis of behavioural patterns in learning activity sequences," in *Proceedings of the 2nd International Conference on Intelligent Networking and Collaborative Systems (INCOS '10)*, pp. 100–107, IEEE, November 2010.
- [6] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.
- [7] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.
- [8] G. David and A. Averbuch, "SpectralCAT: categorical spectral clustering of numerical and nominal data," *Pattern Recognition*, vol. 45, no. 1, pp. 416–433, 2012.
- [9] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, 2007.
- [10] C. Wang, Z. She, and L. Cao, "Coupled attribute analysis on numerical data," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 1736–1742, AAAI Press, August 2013.
- [11] F. Li, G. Xu, L. Cao, X. Fan, and Z. Niu, "CGMF: coupled group-based matrix factorization for recommender system," in *Web Information Systems Engineering—WISE 2013*, vol. 8180 of *Lecture Notes in Computer Science*, pp. 189–198, Springer, Berlin, Germany, 2013.
- [12] A. Jakulin and I. Bratko, *Analyzing Attribute Dependencies*, Springer, Berlin, Germany, 2003.
- [13] T. Calders, B. Goethals, and S. Jaroszewicz, "Mining rank-correlated sets of numerical attributes," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pp. 96–105, ACM, Philadelphia, Pa, USA, August 2006.
- [14] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Relation adaptation: learning to extract novel relations with minimum supervision," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI '11)*, pp. 2205–2210, July 2011.
- [15] K. Niu, W. Chen, Z. Niu et al., "A user evaluation framework for web-based learning systems," in *Proceedings of the 3rd ACM*

International Workshop on Multimedia Technologies for Distance Learning, pp. 25–30, ACM, 2011.

- [16] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: analysis and an algorithm,” in *Advances in Neural Information Processing Systems*, vol. 2, pp. 849–856, 2002.
- [17] S. Boriah, V. Chandola, and V. Kumar, “Similarity measures for categorical data: a comparative evaluation,” *Red*, vol. 30, no. 2, p. 3, 2008.
- [18] S. Cost and S. Salzberg, “A weighted nearest neighbor algorithm for learning with symbolic features,” *Machine Learning*, vol. 10, no. 1, pp. 57–78, 1993.
- [19] K. Niu, Z. Niu, D. Liu, X. Zhao, and P. Gu, “A personalized user evaluation model for web-based learning systems,” in *Proceedings of the IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI '14)*, pp. 210–216, Limassol, Cyprus, November 2014.
- [20] National Academy of Engineering and National Research Council, *Engineering in K-12 Education: Understanding the Status and Improving the Prospects*, National Academies Press, 2009.
- [21] C. Okoli and S. D. Pawlowski, “The Delphi method as a research tool: an example, design considerations and applications,” *Information & Management*, vol. 42, no. 1, pp. 15–29, 2004.
- [22] D.-C. Li and C.-W. Liu, “Extending attribute information for small data set classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 452–464, 2012.
- [23] Y. Jia and C. Zhang, “Instance-level semisupervised multiple instance learning,” in *Proceedings of the 23rd National Conference on Artificial intelligence (AAAI '08)*, pp. 640–645, 2008.