A Coupled User Clustering Algorithm for Web-based Learning Systems

Ke Niu[*†]
kk511@bit.edu.cn

Zhendong Niu[*✉]
zniu@bit.edu.cn

Xiangyu Zhao[*]
koopr@bit.edu.cn

Can Wang[‡]
can.wang@csiro.au

Kai Kang[*]
kangkai@bit.edu.cn

Min Ye[*]
2120141074@bit.edu.cn

[*]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[†]Computer School, Beijing Information Science and Technology University, Beijing, China
[‡]Digital Productivity, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, Australia

ABSTRACT

User clustering algorithms have been introduced to analyze users' learning behaviors and help to provide personalized learning guides in traditional Web-based learning systems. However, the explicit and implicit coupled interactions, which means the correlations between user attributes generated from learning actions, are not considered in these algorithms. Much significant and useful information which can positively affect clustering accuracy is neglected. To solve the above issue, we proposed a coupled user clustering algorithm for Wed-based learning systems. It respectively takes into account intra-coupled and inter-coupled relationships of learning data, and utilizes Taylor-like expansion to represent their integrated coupling correlations. The experiment result demonstrates the outperformance of the algorithm in terms of efficiently capturing correlations of learning data and improving clustering accuracy.

Keywords

Web-based learning, coupled interactions, user clustering, user behavior analysis

1. INTRODUCTION

Information technology and its application have brought great changes to all aspects of human, especially education area. Web-based learning is a significant and advanced way of education, meaning to utilize computer network technology, digital multimedia technology, database technology and other modern information technology to learn in digital environment. Compared with traditional learning, Web-based learning can efficiently meet learners' needs of learning anytime and anywhere. Meanwhile, it takes advantage of various online resources and helps learners to expand their horizons and discover interests.

Recently Web-based learning systems are studied by many education institutions and researchers, and a large number of online learning communities and virtual schools arise [1]. As an emerging online learning system, MOOC (Massive Open Online Courses) was initiated by America's top universities in 2012. It had a participation of more than 6 million of students from around 220 countries within one year [2]. Some of Web-based learning systems apply user clustering algorithms to analyze learning behaviors and provide personalized learning services. Fu and Ofoghlu put forward a new clustering algorithm; it can extract clusters which can be described by overlapping layered concept in dense space [3]. According to the feedback of basic clustering method, Montazer et al. proposed a hybrid clustering algorithm, which considered clustering issues from different perspectives, and kept the simplicity of basic clustering algorithm [4]. Another matrix-based improved clustering algorithm was put forward by Zhang et al., and it is much more efficient when comparing with K-means [5]. Lin et al. came up with a kind of intuitionistic fuzzy kernel clustering algorithm (KIFCM), combining intuitionistic fuzzy sets and fuzzy kernel clustering algorithm, and applied it in learner behavior analysis [6].

With the above algorithms utilized in Wed-based learning systems, learners' attribute information is extracted by analyzing their behaviors, and finally used for user clustering. However, these algorithms generally neglect the explicit and implicit coupling relationships of user attributes and this may lead to massive significant information loss. For example, table 1 presents an evaluation index system based on information provided by a specified Web-based learning system. With common sense, we think that user attribute of "Average correct rate of homework" has a positive impact on "Comprehensive test result". Generally, if the "Average correct rate of homework" is better, the "Comprehensive test result" is better. Students who behave this way are categorized in "normal" group. However, there are also students who can either get a better "Average correct rate of homework" with a worse "Comprehensive test result", or a better "Comprehensive test result" with a worse "Average correct rate of homework"; they are categorized in "unnormal" group. These unnormal situations are caused by irregular correlations of user attributes, but they are often ignored. This will certainly have negative effect on user clustering

Table 1: 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

accuracy.

Nowadays an increasing number of researchers are studying the interactions between object attributes with special attention 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 [7], Wang et al. put forward the intra-coupled and inter-coupled interactions of continuous attributes [8]. An innovative coupled group-based matrix factorization model for discrete attributes of recommender system was addressed by Li et al. [9]. Jakulin and Bratko proposed an algorithm to detect interactions between attributes, but it is only applicable in supervised learning with the experimental results [10]. For unsupervised learning, the coupled nominal similarity to extract new relationships between entities was addressed by Wang et al., but it is only for categorical data [11]. We rarely find any methods applied in Web-based learning systems, that consider coupling relationships of user attributes in user clustering.

This paper proposed a coupled user clustering algorithm for Web-based learning systems, namely CUCA. It studies the coupling relationships of user attributes. With the help of Taylor-like expansion, we use a spectral clustering algorithm to cluster users. When it is applied in Web-based learning systems, it can efficiently capture learners' behavior features and analyze the information behind them, especially that of "unnormal" group of learners, and finally use them to provide personalized learning services. To verify the outperformance of CUCA, we compare its clustering result with that of 3 other algorithms, respectively from 3 dimensions of learning attitude, learning effect and the integrated dimension.

The rest of the paper is organized as following. The clustering algorithm model is proposed in section 2. Section 3 introduces the formalization and exemplification of the clustering algorithm. In section 4, experiments and results analysis are demonstrated. Section 5 concludes the paper and discusses some potential applications of the proposed algorithm in the future.

2. CLUSTERING MODEL

Evaluation model usually plays the core role in user evaluation framework [12]. In this section, the coupled user clustering model is illustrated. This model captures coupling relationships of user attributes through online behavior analysis, and uses spectral clustering algorithm to improve clustering accuracy.

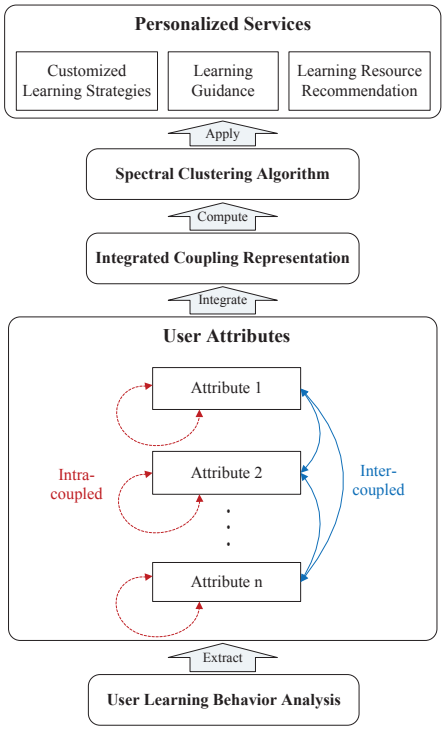


Figure 1: The coupled user clustering model

The model is composed of user learning behavior analysis, coupled interactions computation of user attributes, integrated coupling representation and spectral clustering algorithm, illustrated in figure 1. As the basis, data for user learning behavior analysis needs to be collected, consolidated and normalized. From the data, user attributes information is extracted. With the extracted user attributes, the intra-coupled interaction within an attribute and inter-coupled interaction among different attributes are respectively captured. Then all the interactions are integrated and represented using Taylor-like expansion. Finally we use a spectral clustering algorithm - NJW to cluster users. This model is consequently applied in various Web-based personalized services, like Learning guide customization, tutoring and learning resources recommendation.

3. CLUSTERING ALGORITHM

Based on the model illustrated in section 2, this paper proposed an online coupling user clustering algorithm. It is

Table 2: A fragment example of user 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

suitable for network education, not only applicable to user clustering analysis in Web-based learning systems, but also to enterprise training, performance review and others with users participation and behaviors recording. This section describes the details of the proposed coupled user clustering algorithm. Firstly, it collects user learning behavior information and extracts user attributes from them. Secondly, it calculates and represents users' intra-coupled and inter-coupled relationship. Thirdly, the intra-coupled and inter-coupled interactions are integrated to be a coupled representation. Finally, it clusters users based on the processed attributes, using NJW algorithm.

3.1 User learning behavior analysis

When students login a Web-based learning system, the system will record their activity information, such as number of learning resources, total time length of learning resources and average correct rate of homework, which can be used to build an evaluation index system. We refer to a Web-based personalized user evaluation model [13] and utilizes its evaluation index system to extract students' attributes information. This index system is with evaluation standards of America K-12 (kindergarten through twelfth grade) [14] and Delphi method [15], which is a hierarchical structure built according to mass of information and data generated during general e-learning activities. It defines 20 indicators and can comprehensively represent the students' attributes, as shown in table 1.

Generally attributes are with various data types and units, we formalize them by creating the table 2.

3.2 Intra-coupled and inter-coupled representation

In this section, we represent intra-coupled and inter-coupled interactions of user attributes. And with a few examples, the application of CUCA is demonstrated. We choose 5 students and 6 of the 20 attributes in table 1, which are "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". The 6 attributes are respectively signified by a_1, a_2, a_3, a_4, a_5 and a_6 in table 2.

Here we use a tetrad $S = \langle U, A, V, f\rangle$ to represent user attributes information. $U = \{u_1, u_2, \ldots, u_m\}$ means a finite set of users; $A = \{a_1, a_2, \ldots, a_n\}$ refers to a finite set of attributes; $V = \bigcup_{j=1}^{n} V_j$ represents all attributes value sets; $V_j = \{a_j{\cdot}v_1, \ldots, a_j{\cdot}v_{t_j}\}$ is the value set of the j-th attribute; $f = \bigcup_{i=1}^{n} f_j$, $f_j : U \to V_j$ is the function for calculating a certain attribute value. For example, the information table 2 above 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 $f_1(u_1) = 0.61$.

The common way to calculate the interactions between 2 attributes is Pearson's correlation coefficient [7]. The user attributes from the Table 1 are continuous variables and approximate to Normal distribution, meeting the constraint condition of the Pearson's correlation coefficient. Thus we use it to help to calculate attributes interactions in this paper. For instance, the Pearson's correlation coefficient between a_k and a_j is formalized as:

$$Cor(a_j, a_k) = \frac{\sum_{u\in U}(f_j(u) - \mu_j)(f_k(u) - \mu_k)}{\sqrt{\sum_{u\in U}(f_j(u) - \mu_j)^2 \sum_{u\in U}(f_k(u) - \mu_k)^2}} \tag{1}$$

Where μ_j, μ_k are respectively mean values of a_j, a_k.

The Pearson's correlation coefficient helps to calculate the attributes interactions, but it fits for linear relationship only, which is not sufficient to fully capture pairwise attributes interactions. Therefore we converts the original data attributes into a higher dimensional feature space to extract more attribute information [16].

Firstly, we use a few additional attributes to expand interaction space. Then there are L attributes for each original attribute a_j, including itself, namely $\langle a_j\rangle^1$, $\langle a_j\rangle^2$, ... , $\langle a_j\rangle^L$. Each attribute value is the power of the attribute, for instance, $\langle a_j\rangle^3$ is the third power of attribute a_j, $\langle a_j\rangle^p$ $(1 \le p \le L)$ is the p-th power of a_j. In table 3, the denotation a_j and $\langle a_j\rangle^1$ are equivalent; the value of $\langle a_j\rangle^2$ is the square of that of a_j. For simplicity, we set L=2.

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 the maximum correlation observed by random chance, while the true correlation is zero. If p-value is smaller than 0.05, the correlation $Cor(a_j, a_k)$ is significant. The updated correlation coefficient is as:

$$R\_Cor(a_j, a_k) = \begin{cases} Cor(a_j, a_k) & \text{if } p\text{-value}<0.05, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Here we do not consider all relationships, but only takes the significant coupling relationships into account, because all relationships involvement may cause the over-fitting issue on modeling coupling relationship. This issue will go against the attribute inherent interaction mechanism. So based on the updated correlation, the intra-coupled and inter-coupled interaction of attributes is proposed. Intra-coupled interaction is the relationship between a_j and all its powers; inter-coupled interaction is the relationship between a_j and powers of the rest attributes a_k $(k \ne j)$.

Table 3: Extended user attributes

U \ $\widetilde{A}$	$\langle a_1\rangle^1$	$\langle a_1\rangle^2$	$\langle a_2\rangle^1$	$\langle a_2\rangle^2$	$\langle a_3\rangle^1$	$\langle a_3\rangle^2$	$\langle a_4\rangle^1$	$\langle a_4\rangle^2$	$\langle a_5\rangle^1$	$\langle a_5\rangle^2$	$\langle a_6\rangle^1$	$\langle a_6\rangle^2$
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

Definition 1 Intra-coupled interaction. The intra-coupled interaction within an attribute is represented as a matrix. For attribute a_j, it is an $L \times L$ matrix $R^{Ia}(a_j)$. In the matrix, (p, q) is the correlation between $\langle a_j\rangle^p$ and $\langle a_j\rangle^q$ $(1 \le p, q \le L)$.

$$R^{Ia}(a_j) = \begin{pmatrix} \alpha_{11}(j) & \alpha_{12}(j) & \cdots & \alpha_{1L}(j) \\ \alpha_{21}(j) & \alpha_{22}(j) & \cdots & \alpha_{2L}(j) \\ \cdots & \cdots & \ddots & \cdots \\ \alpha_{L1}(j) & \alpha_{L2}(j) & \cdots & \alpha_{LL}(j) \end{pmatrix} \quad (3)$$

Where $\alpha_{pq}(j) = R\_Cor(\langle a_j\rangle^p, \langle a_j\rangle^q)$ is the Pearson's correlation coefficient between $\langle a_j\rangle^p$ and $\langle a_j\rangle^q$.

For attribute a_1 in table 3 above, we can get the intra-coupled interaction of it as $R^{Ia}(a_1) = \begin{pmatrix} 1 & 0.986 \\ 0.986 & 1 \end{pmatrix}$, 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.

Definition 2 Inter-coupled interaction. The inter-coupled interaction between attribute a_j and other attributes a_k $(k \ne j)$ is quantified as an $L \times L * (n-1)$ matrix as:

$$R^{Ie}(a_j|\{a_k\}_{k\ne j}) = \begin{pmatrix} R^{Ie}(a_j|a_{k_1}) & \cdots & R^{Ie}(a_j|a_{k_{n-1}}) \end{pmatrix} \quad (4)$$

$$R^{Ie}(a_j|a_{k_i}) = \begin{pmatrix} \beta_{11}(j|k_i) & \beta_{12}(j|k_i) & \cdots & \beta_{1L}(j|k_i) \\ \cdots & \cdots & \ddots & \cdots \\ \beta_{L1}(j|k_i) & \beta_{L2}(j|k_i) & \cdots & \beta_{LL}(j|k_i) \end{pmatrix} \quad (5)$$

Here $\{a_k\}_{k\ne j}$ refers to all the attributes except for a_j, and $\beta_{pq}(j|k_i) = R\_Cor(\langle a_j\rangle^p, \langle a_{k_i}\rangle^q)$ is the correlation coefficient between $\langle a_j\rangle^p$ and $\langle a_{k_i}\rangle^q$ $(1 \le p, q \le L)$.

For attribute a_1 in the table 3 above, the inter-coupled interaction between a_1 and others $(a_2, a_3, a_4, a_5, a_6)$ is calculated as:

$$R^{Ie}(a_1|\{a_2, a_3, a_4, a_5, a_6\}) =$$

$$\begin{pmatrix} 0 & 0 & 0.898 & 0.885 & 0.928 & 0.921 \\ 0 & 0 & 0.929 & 0.920 & 0.879 & 0.888 \end{pmatrix}$$

$$\begin{pmatrix} 0.997 & 0.982 & 0.999 & 0.988 \\ 0.978 & 0.994 & 0.982 & 0.999 \end{pmatrix}$$

The p-values between a_1 and others $(a_2, a_3, a_4, a_5, a_6)$ is calculated as:

$$p^{Ie}(a_1|\{a_2, a_3, a_4, a_5, a_6\}) =$$

$$\begin{pmatrix} 0.689 & 0.677 & 0.039 & 0.046 & 0.023 & 0.027 \\ 0.733 & 0.707 & 0.023 & 0.027 & 0.050 & 0.044 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0.003 & 0 & 0.002 \\ 0.004 & 0.001 & 0.003 & 0 \end{pmatrix}$$

Based on the result, we can find that there is hidden correlation between user attributes. For instance, all the p-values between attribute a_1 and a_2 are larger than 0.05, so the correlation coefficient is 0 based on Equation (2), 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, a_1 and a_6 is quite close to 1; it indicates "Daily average quiz result" and "Comprehensive test result" respectively have close relationship with "Average correct rate of homework", which is consistent with our practical experiences. In conclusion, comprehensively taking into account intra-coupled and inter-coupled correlation of attributes can efficiently help capturing coupling relationships between user attributes.

3.3 Integrated coupling representation

Intra-coupled and inter-coupled interactions are integrated in this section as a coupled representation scheme.

In table 3 above, each user is signified by $L*n$ updated variables $\widetilde{A} = \{\langle a_1\rangle^1, \ldots, \langle a_1\rangle^L, \ldots, \langle a_n\rangle^1, \ldots, \langle a_n\rangle^L\}$. With the updated function $\widetilde{f}_j^p(u)$, the corresponding value of attribute $\langle a_n\rangle^p$ is assigned to user u. Attribute a_j and all its powers are signified as $\widetilde{u}(a_j) = [\widetilde{f}_j^1(u), \ldots, \widetilde{f}_j^L(u)]$, while the rest attributes and all powers are presented in another vector $\widetilde{u}(\{a_k\}_{k\ne j}) = [\widetilde{f}_{k_1}^1(u), \ldots, \widetilde{f}_{k_1}^L(u), \ldots, \widetilde{f}_{k_{n-1}}^1(u), \ldots, \widetilde{f}_{k_{n-1}}^L(u)]$. For instance, in table 3, $\widetilde{u_1}(a_1) = [0.61, 0.37]$, $\widetilde{u_1}(\{a_2, a_3, a_4, a_5, a_6\}) = [0.55, 0.30, 0.47, 0.22, 0.72, 0.52, 0.63, 0.40, 0.62, 0.38]$.

Definition 3 Coupled representation. Attribute a_j's coupled representation is formalized as a $1\times L$ vector $u^c(a_j|\widetilde{A}, L)$,

where $(1, p)$ component corresponds to the updated attribute $\langle a_j \rangle^p$.

$$u^c(a_j|\widetilde{A}, L) = u^{Ia}(a_j|\widetilde{A}, L) + u^{Ie}(a_j|\widetilde{A}, L) \tag{6}$$

$$u^{Ia}(a_j|\widetilde{A}, L) = \widetilde{u}(a_j) \odot w \otimes [R^{Ia}(a_j)]^T \tag{7}$$

$$\begin{aligned} u^{Ie}(a_j|\widetilde{A}, L) = \widetilde{u}(\{a_k\}_{k\neq j}) \odot [w, w, \ldots, w] \\ \otimes [R^{Ie}(a_j|\{a_k\}_{k\neq j})]^T \end{aligned} \tag{8}$$

where $w = [1/(1!), 1/(2!), \ldots, 1/(L!)]$ is a constant $1 \times L$ vector, $[w, w, \ldots, w]$ 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.

Take an example in table 4, the coupled representation for attribute a_1 is presented as $u_1^c(a_1|\widetilde{A}, 2) = [3.85, 3.80]$. The reason we choose such a representation method is explained below. If the above Equation (6) 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 [17].

$$\begin{aligned} u^c(a_j|\widetilde{A}, L) \cdot \langle a_j \rangle^p = \alpha_{p1}(j) \cdot \widetilde{f}_j^1(u) + \sum_{i=1}^{n-1} \frac{\beta_{p1}(j|k_i)}{1!} \widetilde{f}_{k_i}^1(u) \\ + \frac{\alpha_{p2}(j)}{2!} \widetilde{f}_j^2(u) + \sum_{i=1}^{n-1} \frac{\beta_{p2}(j|k_i)}{2!} \widetilde{f}_{k_i}^2(u) + \ldots \\ + \frac{\alpha_{pL}(j)}{L!} \widetilde{f}_j^L(u) + \sum_{i=1}^{n-1} \frac{\beta_{pL}(j|k_i)}{L!} \widetilde{f}_{k_i}^L(u) \end{aligned} \tag{9}$$

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

$$u^c(\widetilde{A}, L) = [u^c(a_1|\widetilde{A}, L), u^c(a_2|\widetilde{A}, L), \ldots, u^c(a_n|\widetilde{A}, L)] \tag{10}$$

With the couplings of attributes, each user is represented as a $1 \times L*n$ vector. When all the users follow the steps above, we then obtain an $m \times L*n$ coupled information table. For example, based on table 2, the coupled information table shown in table 4, is the new representation.

3.4 User clustering

We obtained the global coupled representation in table 4. Compared with the original representation, this one reflects coupling interactions of attributes, and contains far more coupling relationships. With these data, we can do user clustering using NJW [18], which is a kind of spectral clustering algorithm. Detailed clustering results are demonstrated in experiment later.

4. EXPERIMENTS AND EVALUATION

In this section, we conduct experiments to verify the validity and accuracy of the proposed algorithm. 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"[1]. As a basic platform for national lifelong education, which started the earliest in China, and had the largest group of users and provided most extensive learning resources, it met the needs of personalization and user diversity through integrating a variety of multi-network, terminals and resources. So far, the number of registered users has reached more than two million. The experiment is composed of 3 parts: user study, user clustering and result analysis.

4.1 User study

In the experiment, we ask 220 users (signified by $s_1, s_2, \ldots, s_{220}$) to learn C programming language online. The whole learning process, including recording and analyzing learning activities information, is accomplished in CETV.

The public data sets regarding learners' learning behaviors in online learning systems are insufficient, and most of them don't contain labeled user clustering information. Meanwhile, because learners always behave with certain subjectivity in online learning systems, to label learners with different classifiers based on their learning behaviors only, but without the information behind, is not accurate. Therefore, we adopt a few user study methods, including self-assessment, peer-assessment and teacher-assessment [19], to label online learners with classifiers. It is the basis for verifying the accuracy of clustering.

Analyzing the 20 attributes extracted from table 1 using user evaluation index system proposed in this paper, we can easily find that they can be mainly divided into 2 categories. Some attributes belong to the category of "learning attitude", which refers to students' learning initiatives, like "Times of doing homework", "Number of learning resources" and "Total time length of learning resources". While the rest belong to the category of "learning effect", which refers to how well students receive knowledge, like "Average correct rate of homework", "Daily average quiz result" and "Comprehensive test result". Accordingly, we can label learners with these attributes from both categories. Each of the attributes has 3 grades - high, medium and low. Consequently every learner has 2 labels and each label has one grade of high, medium and low. In total, there will be 9 different combinations - high & high, high & medium, high & low, medium & high, medium & medium, medium & low, low & high, low & medium and low & low.

After the students had finished a learning phase, we asked the 220 users to do a self-assessment using centesimal grade, respectively from perspectives of learning attitude and learning effect. Then we requested teacher assessments in the same way, meaning the teacher of the subject to review the students' performance. Finally, the students were asked to do peer-assessments, which means students do an assessment for each other. Each student will get the assessment scores from the rest 219 students. We calculate the aver-

[1]http://www.guoshi.com/

Table 4: Integrated coupling representation of user attributes

U \ $\widetilde{A}$	$\langle a_1\rangle^1$	$\langle a_1\rangle^2$	$\langle a_2\rangle^1$	$\langle a_2\rangle^2$	$\langle a_3\rangle^1$	$\langle a_3\rangle^2$	$\langle a_4\rangle^1$	$\langle a_4\rangle^2$	$\langle a_5\rangle^1$	$\langle a_5\rangle^2$	$\langle a_6\rangle^1$	$\langle a_6\rangle^2$
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

Table 5: Transformation rule between score and grade

Score range	Grade	Sample
$80 \leq X \leq 100$	high	95
$50 \leq X < 80$	medium	75
$0 \leq X < 50$	low	40

Table 6: The evaluation results of s_1

	learning attitude	learning effect
Self-assessment (40%)	80.0	75.0
Teacher-assessment (35%)	85.0	80.0
Peer-assessment (25%)	82.7	79.2
Comprehensive evaluation results	82.4	77.8
grade	high	medium
Class	high & medium	

age of the 219 scores. A student's final score is obtained by integrating the 3 assessments above. According to Expert Investigation Weight Method [15], we did statistical analysis and got approximate weights for the assessments, namely 40% for self-assessment, 35% for teacher-assessment and 25% for peer-assessment. Each student's final score will be transformed into a grade value, "high", "medium" or "low". The transformation rule between score and grade is shown in table 5.

Take student s_1 as an example, his 3 assessment scores and transformed grades are shown in table 6.

4.2 User clustering

In Equation (9), the proposed coupled representation is strongly dependent on how large L can be. Thus we conduct a few experiments to study how the performance of L influences the clustering accuracy of CUCA. The range of L value is from $L = 1$ to $L = 10$. With the growth of L value, $L!$ value grows. When L = 10, it is large enough to capture most of the information in Equation (9). The experiments show that with the growth of L, the clustering accuracy will be gradually improved. When $L = 3$, the accuracy change reaches a comparatively stable status; when $L > 3$, the accuracy change is extremely small. That means the accuracy of when $L = 3$ and when $L = 10$ is quite similar. To guarantee the accuracy of experimental results and reduce the complexity of the algorithm, we take $L = 3$ in the following comparative experiments.

In the experiments, we utilize the attributes data generated from the 220 students' learning process, as the basis for clustering. Then we persistently collect data from the process which reaches 30 hours by average. Respectively with the help of K-means algorithm, Fuzzy C-means algorithm(FCM), NJW algorithm and CUCA algorithm, we do user clustering, getting 2 labels in terms of learning attitude and learning effect for each student. In section 4.1, we classified each student with 2 labels based on user study result. Then we compare the labels got from user study and user clustering result. If only one label from each side is the same, the clustering accuracy rate is 50%; if both the labels are the same, the accuracy rate reaches 100%. For instance, student s_1 is labeled with "high & medium" in user study, if he is classified to "medium & medium" cluster, the clustering accuracy rate is 50%; if he is classified to "high & medium" cluster, the accuracy rate reaches 100%.

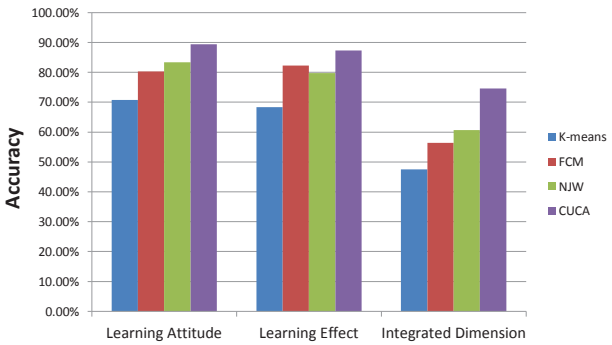
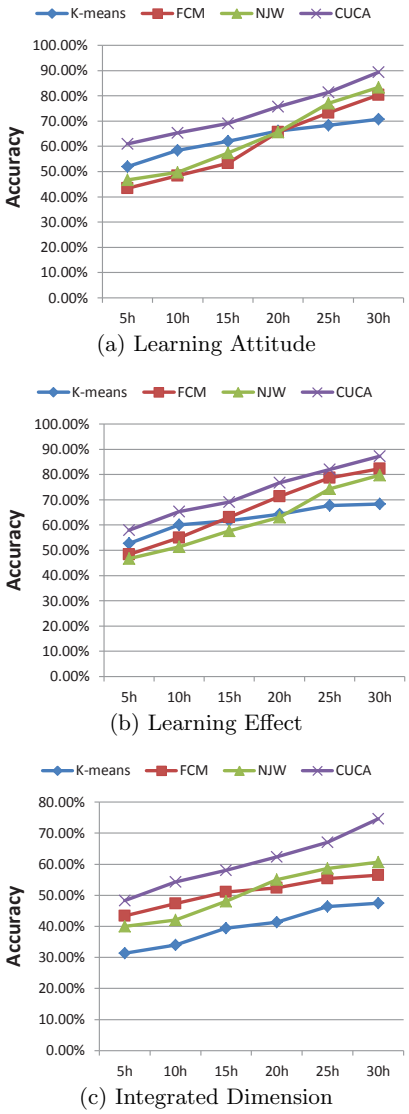


Figure 2: Clustering result analysis (30h)

4.3 Result analysis

We do comparison analysis on the clustering result respectively from the 3 dimensions of learning attitude, learning effect and the integrated dimension. The analysis result is shown in figure 2. We can see the clustering accuracy of utilizing CUCA is 89.4% for learning attitude, 87.3% for learning effect and 74.6% for integrated dimension, each of which is higher than that with the other 3 algorithms. Especially, CUCA obviously outperforms the rest on clustering accuracy of integrated dimensions. Compared with K-means, which performs the worst, CUCA improves almost 30% on

the clustering accuracy. The reason is CUCA fully takes into account coupling relationships of users. In Web-based learning systems, if the user attributes are more complicated, there will be more clustering dimensions and the clustering accuracy will be improved more.



(a) Learning Attitude

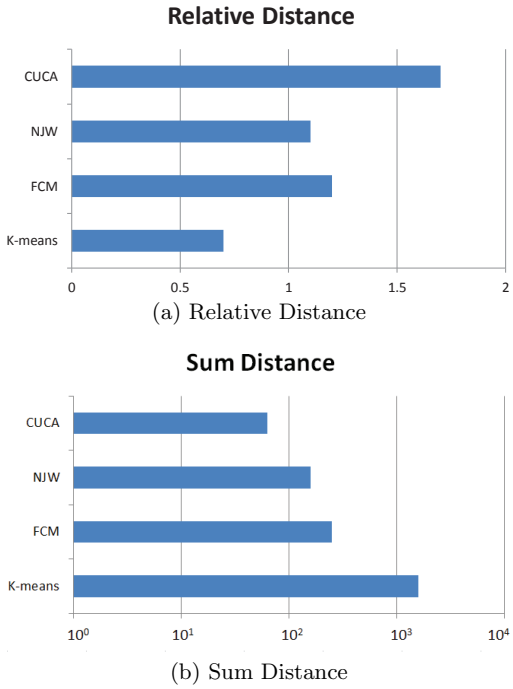
(b) Learning Effect

(c) Integrated Dimension

Figure 3: Clustering result of different time phases

If we divide the process of extracting user attributes to 6 phases, namely 5h, 10h, 15h, 20h, 25h, 30h based on average learning length, we can get the correlation between average learning length and clustering accuracy, as shown in figure 3. From the figure, we can see that while the learning length grows, the clustering accuracy of the 4 algorithms keeps improving, specifically for CUCA. With CUCA, the clustering accuracy on integrated dimensions distinctly outperforms that of the 3 other algorithms. It indicates that with the increasing learning behavior data volume, CUCA can find the hidden coupling relationships of user attributes more easily, and the clustering accuracy is much better.

Besides, we can verify clustering accuracy through analyzing user clustering results. The best performance of a clustering algorithm is keeping the distance within clusters as small as possible and the distance between clusters as large as possible. We use the evaluation criteria of Relative Distance (the ratio of average inter-cluster distance upon average intra-cluster 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 4, we can see that the Relative Distance for CUCA is larger than that of the 3 other algorithms, while the Sum Distance for CUCA is smaller. It indicates that CUCA outperforms the rest in terms of clustering structure.



(a) Relative Distance

(b) Sum Distance

Figure 4: Clustering structure analysis (30h)

5. CONCLUSION

A coupled user clustering algorithm (CUCA) for Web-based learning systems is proposed in this paper to capture coupling relationships of user attributes. The algorithm respectively takes intra-coupled and inter-coupled correlation into account in the application process, and utilizes Taylor-like expansion to represent the coupling relationship. Finally, with the usage of spectral clustering algorithm, CUCA is applied to do user clustering. In the experiments, user study, user clustering and result analysis are adopted to verify that CUCA outperforms traditional algorithm for user clustering.

In this paper, the user attributes extracted from user learning behavior data are all numerical data, most of which are continuous data. In reality, there are also categorical data, which will be a significant study topic in the future.

6. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Project No. 61370137), the National "973" Project of China (No. 2012CB720702) and Major Science and Technology Project of Press and Publication (No: GAPP_ZDKJ_BQ/01).

7. 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." *iJIM*, vol. 8, no. 1, pp. 46–50, 2014.

[3] H. Fu and M. Ó. FoghlÙ, "A conceptual subspace clustering algorithm in e-learning," in *Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on*, vol. 3. IEEE, 2008, pp. 1983–1988.

[4] G. A. Montazer and M. S. Rezaei, "A new approach in e-learners grouping using hybrid clustering method," in *Education and e-Learning Innovations (ICEELI), 2012 International Conference on*. IEEE, 2012, pp. 1–5.

[5] K. Zhang, L. Cui, H. Wang, and Q. Sui, "An improvement of matrix-based clustering method for grouping learners in e-learning," in *Computer Supported Cooperative Work in Design, 2007. CSCWD 2007. 11th International Conference on*. IEEE, 2007, pp. 1010–1015.

[6] K. Lin, C. Lin, K. Hung, Y. Lu, and P. Pai, "Developing kernel intuitionistic fuzzy c-means clustering for e-learning customer analysis," in *Industrial Engineering and Engineering Management (IEEM), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1603–1607.

[7] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Siam, 2007, vol. 20.

[8] C. Wang, Z. She, and L. Cao, "Coupled attribute analysis on numerical data." in *IJCAI*, 2013.

[9] 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*. Springer, 2013, pp. 189–198.

[10] A. Jakulin and I. Bratko, *Analyzing attribute dependencies*. Springer, 2003.

[11] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou, "Coupled nominal similarity in unsupervised learning," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 973–978.

[12] K. Niu, W. Chen, Z. Niu, P. Gu, Y. Li, and Z. Huang, "A user evaluation framework for web-based learning systems," in *Proceedings of the third international ACM workshop on Multimedia technologies for distance learning*. ACM, 2011, pp. 25–30.

[13] K. Niu, Z. Niu, D. Liu, X. Zhao, and P. Gu, "A personalized user evaluation model for web-based learning systems," in *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*. IEEE, 2014, pp. 210–216.

[14] L. Katehi, G. Pearson, and M. Feder, *Engineering in K-12 Education: Understanding the Status and Improving the Prospects*. National Academies Press, 2009.

[15] 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.

[16] D. Li and C. Liu, "Extending attribute information for small data set classification," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 3, pp. 452–464, 2012.

[17] Y. Jia and C. Zhang, "Instance-level semisupervised multiple instance learning." in *AAAI*, 2008, pp. 640–645.

[18] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[19] C. Chang, K. Tseng, and S. Lou, "A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a web-based portfolio assessment environment for high school students," *Computers & Education*, vol. 58, no. 1, pp. 303–320, 2012.