A Systematic Review on Educational Data Mining

Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan

***Abstract*— Presently educational institutions compile and store huge volumes of data such as student enrolment and attendance records, as well as their examination results. Mining such data yields stimulating information that serves its handlers well. Rapid growth in educational data points to the fact that distilling massive amounts of data requires a more sophisticated set of algorithms. This issue led to the emergence of the field of Educational Data Mining (EDM). Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. One such preprocessing algorithm in EDM is Clustering. Many studies on EDM have focused on the application of various data mining algorithms to educational attributes. Therefore, this paper provides over three decades long (1983-2016) systematic literature review on clustering algorithm and its applicability and usability in the context of EDM. Future insights are outlined based on the literature reviewed, and avenues for further research are identified.**

***Index Terms*— Data mining; Clustering methods; Educational technology; Systematic Review.**

I. INTRODUCTION

AS an interdisciplinary field of study, Educational Data Mining (EDM) applies machine-learning, statistics, Data Mining (DM), psycho-pedagogy, information retrieval, cognitive psychology, and recommender systems methods and techniques to various educational data sets so as to resolve educational issues [1]. The International Educational Data Mining Society [2] defines EDM as "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (p. 601). EDM is concerned with analyzing data generated in an educational setup using disparate systems. Its aim is to develop models to improve learning experience and institutional effectiveness. While DM, also referred to as Knowledge Discovery in Databases (KDDs), is a known field of study in life sciences and commerce, yet, the application of DM to educational context is limited [3].

One of the pre-processing algorithms of EDM is known as Clustering. It is an unsupervised approach for analyzing data in statistics, machine learning, pattern recognition, DM, and bioinformatics. It refers to collecting similar objects together to form a group or cluster. Each cluster contains objects that are similar to each other but dissimilar to the objects of other groups. This approach when applied to analyze the dataset derived from educational system is termed as Educational Data Clustering (EDC). An educational institution environment broadly involves three types of actors namely teacher, student and the environment. Interaction between these three actors generates voluminous data that can systematically be clustered to mine invaluable information. Data clustering enables academicians to predict student performance, associate learning styles of different learner types and their behaviors and collectively improve upon institutional performance. Researchers, in the past have conducted studies on educational datasets and have been able to cluster students based on academic performance in examinations [4,5].

Various methods have been proposed, applied and tested in the field of EDM. It is argued that these generic methods or algorithms are not suitable to be applied to this emerging discipline. It is proposed that EDM methods must be different from the standard DM methods due to the hierarchical and non-independent nature of educational data [6]. Educational institutions are increasingly being held accountable for the academic success of their students [7]. Notable research in student retention and attrition rates has been conducted by Luan [8]. For instance, Lin [9] applied predictive modeling technique to enhance student retention efforts. There exist various software's like Weka, Rapid Miner, etc. that apply a combination of DM algorithms to help researchers and stakeholders find answers to specific problems.

The e-commerce websites use recommender systems to collect user browsing data to recommend similar products. There have been efforts to apply the same strategy in the educational information system. One such successful system is the degree compass. [10] a course recommendation system developed by Austin Peay State University, Tennessee. It uses predictive analytical algorithms based on grade and enrollment data to rank courses. Such courses if taken by the student helps them excel through their program of study.

This paper was submitted on 07-Dec-2016 and was accepted on 24-Dec-2016. This research is supported by Universitas Teknologi Yogyakarta research grant 07/UTY-R/SK/0/X/2013.

A. Dutt, M. A. Ismail and T. Herawan* are with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Pantai Valley, Kuala Lumpur, Malaysia. (E-mails: ashishdutt@yahoo.com.my, maizatul@um.edu.my, tutut@um.edu.my)

Tutut Herawan is also with the Universitas Teknologi Yogyakarta and AMCS Research Center, Indonesia. (E-mail: tutut@uty.ac.id)

*Corresponding author, Tel: +603 7967 2509

We have conducted a comprehensive systematic literature review covering research of over three decades (1983-2016) on the applications of clustering algorithms in the educational domain. This is our contribution. Future insights are outlined based on the literature reviewed, and avenues for further research are identified.

This paper is organized as follows. Section II introduces and discusses EDM. Section III provides an introduction to clustering methods. Section IV provides a tabulated format of all the research works that have been carried out till date in EDM using clustering methods. It then continues to provide an analytical discourse on the application of clustering on various educational data-types. Section V discusses the findings; Section VI gives useful insights into the literature gap that was found during the review process and leads to the future course of research. Finally Section VII provides the conclusion.

II. EDUCATIONAL DATA MINING (EDM)

The EDM process converts raw data coming from educational systems into useful information that could potentially have a greater impact on educational research and practice" [1]. Traditionally, researchers applied DM methods like clustering, classification, association rule mining, and text mining to educational context [11]. A survey conducted in 2007, provided a comprehensive resource of papers published between 1995 and 2005 on EDM by Romero & Ventura [12]. This survey covers the application of DM from traditional educational institutions to web-based learning management system and intelligently adaptive educational hypermedia systems.

In another prominent EDM survey by Peña-Ayala [13], about 240 EDM sample works published between 2010 and 2013 were analyzed. One of the key findings of this survey was that most of the EDM research works focused on three kinds of educational systems, namely, educational tasks, methods, and algorithms. Application of DM techniques to study on-line courses was suggested by Zaıane & Luo [14]. They proposed a non-parametric clustering technique to mine offline web activity data of learners. Application of association rules and clustering to support collaborative filtering for the development of more sensitive and effective e-learning systems was studied by Zaıane [15]. The researchers Baker, Corbett & Wagner [16] conducted a case study and used prediction methods in scientific study to game the interactive learning environment by exploiting the properties of the system rather than learning the system. Similarly, Brusilovsky & Peylo [17] provided tools that can be used to support EDM. In their study Beck & Woolf [18] showed how EDM prediction methods can be used to develop student models. It must be noted that student modeling is an emerging research discipline in the field of EDM [6]. While another group of researchers, Garcia at al [19] devised a toolkit that operates within the course management systems and is able to provide extracted mined information to non-expert users. DM techniques have been used to create dynamic learning exercises based on students' progress through English language instruction course by Wang & Liao [20]. Although most of the e-learning systems utilized by educational institutions are used to post or access course materials, they do not provide educators with necessary tools that could thoroughly track and evaluate all the activities performed by their learners to evaluate the effectiveness of the course and learning process [21].

III. CLUSTERING ALGORITHMS

Clustering simply means collecting and presenting similar data items. But what defines similarity? That is the key to understanding 'clustering'. A cluster is therefore a group of items that are similar to each other within the group and dissimilar to objects belonging to other clusters. In statistical notation, "clustering is the most important unsupervised learning algorithm" [22]. Being a pre-processing algorithm in the data mining process, clustering can significantly reduce the data size to meaningful clusters that can be used for further data analysis. However, one must be careful when reducing the data size because when representing data in the form of fewer clusters typically loses certain fine details similar to lossy data compression.

The classification of clustering algorithms is imprecise because several of them overlap with each other. In traditional terms, clustering techniques have broadly been classified into two types, *hierarchical* and *partitional.* But before we discuss these types it's important to understand the subtle difference between clustering (the unsupervised classification) and supervised classification (or discriminant analysis). In supervised classification, we are given a collection of labelled (or pre-classified) data patterns. The objective is to determine the labeling for a newly encountered unlabeled dataset. Whereas, in the case of clustering the problem is to group the unlabeled dataset into meaningful categorical labeled patterns or clusters.

When classifying clustering methods, on the one hand, the nature of the clustering method has to be considered. Thus, concerning the structure of clusters that form the clustering solution (one-layer or several layers of clusters), Partitional and Hierarchical methods are usually distinguished. Furthermore, the distinction between Hard and Soft methods, which is referred to how the objects in the dataset are mapped onto the clusters (binary mapping vs. degree of belonging), is very relevant as well.

While, clustering methods are typically classified according to the approach adopted to implement the algorithm: Centroid-based clustering, Graph-based clustering, Grid-based clustering, Density-based clustering, neural networks-based clustering, and etc. Thus, one can find algorithms that implement Partitional/Hierarchical and hard/soft methods within each and every of these approaches.

Thus, considering these definitions, we can find, for instance: *K*-means/Fuzzy c-means, which are the most typical examples of centroid-based hard/soft partitional clustering algorithm, respectively; Single Link (SLINK) or nearest neighbor is a popular graph-based hard hierarchical clustering algorithm; or Density Based Spatial Clustering of Applications with Noise (DBSCAN), is a density-based hard partitional clustering algorithm.

Continuing further, only geometric hierarchical methods (e.g. Ward's method) consider the existence of cluster centroids when implementing the linkage function, but graph hierarchical methods (such as Single-Link and Complete-Link) are not based on centroids, or any other kind of center-based approach to clustering. Divisive clustering is the other type of hierarchical algorithm. It's the "top-down" approach in which initially all the data points are in one big cluster and splits are performed recursively as the algorithm moves down the hierarchy. Further details on these algorithms can be found in the work of Jain & Dubes [23].

Clustering algorithms are also applied to voluminous data sizes such as big data. The concept of big data refers to voluminous, enormous quantities of data both in digital and physical formats that can be stored in miscellaneous repositories such as records of students' tests or examinations as well as bookkeeping records by Sagiroglu & Sinanc [24]. A data set whose computational size exceeds the processing limit of the software, can be categorized as big data as proposed by Manyika et at [25]. Several studies have been conducted in the past that provide detailed insights into the application of traditional data mining algorithms like clustering, prediction, and association to tame the sheer voluminous power of big data by Zaiane & Luo [14]. Broadly, educational system can be classified as two types; brick or mortar based traditional classrooms and digital virtual classrooms better known as Learning Management Systems (LMSs), web-based adaptive hypermedia systems [26] and intelligent tutoring systems (ITSs) [6].

IV. LITERATURE SEARCH PROCEDURE AND CRITERIA

Since this is a review paper so it becomes important to outline the literature search criteria and the underlying process involved. This study followed Kitchenham, *et al*. [27] methodological guidelines for conducting a systematic literature review. The research question for this study is to agglomerate the application of clustering algorithms to educational data. The major steps for conducting the literature search are as follows;

A. Constructing Search Terms

The following details will help in defining the search terms that we used for our research question. Educational attributes: learning styles, exam failure, classroom decoration, annotation, exam scheduling or timetabling, e-learning, learning outcome, learning objectives, student seating arrangement, student motivation, student profiling, intelligent tutoring systems (ITS), semantic web in education, classroom learning, collaborative learning, education affordability. Clustering algorithms: broadly classified as partition, hierarchical, density, grid type, hard and soft clustering. An example of research question containing the above detail is: [How is *K*-means applied to] CLUSTERING ALGORITHM [learning styles of student] Educational attribute.

B. Search Strategy

We constructed the search terms by identifying the educational attribute and clustering algorithm. We also searched for alternative synonyms, keywords. We used Boolean operators like AND, OR, NOT in our search strings. Five databases were used to search and filter out the relevant papers. The five databases are given in Table I.

TABLE I
DATA SOURCES AND RESULTS FOR LITERATURE SEARCH

No	Data source	Total results	Primary selection	Final selection
a	IEEEXplore	421	152	58
b	ACM Digital	321	143	38
c	JEDM (Journal of Educational Data Mining)	19	19	7
d	ProQuest Education Journals (ABI/Inform Complete, ERIC, ProQuestEducation Journals)	552	80	27
e	ScienceDirect	316	57	36

C. Publication Selection

a. Inclusion Criteria

The inclusion criteria to determine relevant literature (journal papers & magazines, conference papers, technical reports, book's and e-book's, early access articles, standards, education and learning) are listed below:

- Studies that have reviewed educational attribute's in context to clustering approach.
- Studies that analyze educational attributes in context to clustering as a data mining approach.

b. Exclusion Criteria

The following criteria used to exclude literature that was not relevant for this study.

- Studies that are not relevant to the research question.
- Studies that do not describe or analyze the interrelationship between educational data attributes and clustering algorithms.

c. Selecting Primary Sources

The planned selection process for this study had two parts: an initial selection of published papers that could plausibly satisfy the search strings or the selection criteria based on reading the title, abstract and keywords followed by the final selection based on the initially selected list of papers on reading the full text of the paper. The selection process was performed by the primary reviewer. However, to mitigate the primary reviewer's bias if any an inter-rater reliability test was performed in which a secondary reviewer confirmed the primary reviewers result by randomly selecting the set of primary sources (i.e. 15 articles). We have identified 166 articles as our final selection for this review process that are shown in Tables I and II, respectively.

d. Range of Research Papers

The literature review performed in the present study covers published research from year 1983 to year 2016.

V. EDUCATIONAL DATA AND CLUSTERING METHODS

As mentioned in passing, EDC is based on data mining techniques and algorithms and is aimed at exploring educational data to find predictions and patterns in data that characterize learners' behavior. In Table II, we have provided a brief outline of major EDM works that have predominantly applied clustering approach to educational data sets.

TABLE II
CLUSTERING ALGORITHMS AS APPLIED IN EDM

No	Reference	Problem/ Objective	Algorithm/Method	Dataset/Data source	Group
1	[28]	To automatically detect the web usage patterns of users.	SAS FASTCLUS algorithm Ward's algorithm were used	User session transaction logs of the University of California's MELVYL on-line library catalog system.	Clustering learning style
2	[29]	To classify participant learning style	Two-step cluster analysis algorithm of SPSS16 & Microsoft Excel.	80 first year students from Sultan Idris Education University, Malaysia.	Clustering learning style
3	[30]	To determine the optimal parameters and partitions for clustering algorithms	*K*-means, Farthest First &EM in Weka	265 records e-Personalized English Learning System of Xi'an Jiaotong University.	Clustering in e-learning
4	[31]	To derive social-network graphs in student e-learning activities	Hierarchical clustering	Moodle data from LMS of Silesian University	Clustering in e-learning
5	[32]	To provide personalized e-learning environment on learner personality	Fuzzy C means, *K*-means	Data set from Xi'an Jiaotong University"	Clustering in e-learning
6	[33]	To classify the e-learning behaviour of learners	Fuzzy clustering	A non-statitical method to analyze the e-learning behaviour is proposed.	Clustering in e-learning
7	[34]	To recommend the best course combination to students	*K*-means clustering, Apriori association rule	Experiments prove that *K*-means & Apriori association when combined give dense clusters & more associations as compared to only Apriori association rule.	Clustering in e-learning
8	[35]	To cluster the e-learning behaviour of learners	Ward's clustering and *K*-means	Sample size was of 59 students from a Mid-Western University	Clustering in e-learning
9	[36]	To model learner's participation profile in online discussion forums	Hierarchical clustering	672 learners from 18 e-learning classrooms in a degree course from February 2009 to July 2010.	Clustering in e-learning
10	[37]	To determine the influence of human characteristics on user preferences while using WBeI	*K*-means clustering	82 students, expert computer users favored multi-page dynamic buttons and drop-down menus as compared to novice	Clustering in e-learning
11	[38]	To group learners based on their cognitive styles of learning	*K*-means, C-means, evolutionary fuzzy clustering	98 undergraduate students in an e-learning computer networks course	Clustering in e-learning
12	[39]	Analyze the weblog data of Learning Management System	Markov Clustering, Simple *K*-Means	1199 students from Technology Education Institute (TEI) of Kevala using Open eClass (GUNet, 2009)	Clustering in e-learning
13	[40]	To determine the selection of instances and attributes gathered that affect the accuracy and comprehensibility of prediction.	EM, Hierarchical Cluster, SIB, *K*-Means algorithm.	114 university students in first-year course in computer science.	Clustering in e-learning
14	[41]	To identify learning performance assessment rules	Gray relational theory, *K*-means and fuzzy inference	Experimental results indicate that teachers easily assess the learning performance by utilizing only the learning portfolio.	Clustering in e-learning
15	[42]	To group students with similar learning caliber	*K*-means clustering	Sample size was of 70 students	Clustering collaborative learning
16	[43]	To find active & passive collaborators within the group	EM algorithm from Weka	Over 100 student's data from UNED European universities.	Clustering collaborative learning
17	[44]	To determine the key factors essential to the success of educational training	Cluster Analysis, Linkage Method	Personnel educational training database of China Motor Corporation.	Clustering collaborative learning
18	[45]	To identify student's learning ability in a collaborative-learning	Item-Response theory & *K*-means clustering.	116 students participated in the experiment. Average learning ability improved from	Clustering collaborative

		environment		3.84 to 5.97 only improved in control group from 2.16 to 2.4	learning
19	[46]	To evaluate undergraduate performance in semester exam	ANN, Farthest First, Decision Tree	Student data of department of computer science, NUDM.	Clustering in EDM
20	[47]	To identify variables influencing performance of undergraduate students	C-means clustering	Academic database of the Industrial University of Santander (IUS).	Clustering in EDM
21	[48]	To identify student performance from mining historical student record.	CHAID Classification algorithm	A total of 2,228 exam records of foundation students of the UTN, Malaysia. Results show a 70.17 % accuracy of correct prediction.	Clustering exam failure
23	[49]	To investigate the design choices made by teachers in decorating classroom walls enhance learning	*K*-means clustering	30 classrooms of local charter North-Eastern US were studied. Findings suggest that teachers systematically choose to decorate classroom walls.	Clustering in classroom decoration
24	[50]	Students seating choice in classroom & its implications on their assessments	*K*-means clustering	220 students of semester 2011-2012 from University of Novi Sad. Experiments show students with spatial deployment choices scored 10% better than those without it.	Clustering learner seating arrangement
25	[51]	To cluster student's e-learning performance	*K*-means, farthest-first, EM	162 students from Chung Yuan Christian University, Taiwan. Data was taken from i-learning [52]	Clustering learning portfolio
26	[53]	To modeling student e-learning behavior for effective and adaptive teaching are studied	C4.5 algorithm and Bayesian Markov Chain	Data set of 89 student's interactions with AToL (Adaptive Tutor for online Learning) by taking the CS-1 course in 2005.	Clustering Student modeling
27	[54]	To measure the shallowness of student learning	Statistical measures	71 undergraduate students using the Genetics Cognitive tutor enrolled in genetics classes at Carnegie Mellon University.	Clustering Student modeling
28	[55]	To recommend webpages to student based on their web surfing behaviour.	Hierarchical *K*-means clustering	42,633 webpages collected in 30 days from a computer lab and segmented into 19 clusters.	Student profiling clustering
29	[56]	To discover student profiles from course evaluation data and for generating associations between subjects based on the student performance.	EM, association-rule and decision tree	The data was collected from Polytechnic University of Tirana (UPT) in three different bachelor programs.	Student profiling clustering
30	[57]	To improve graduate students' performance, and overcome the problem of low grades.	Lift-metric, Rule-based & Naïve Bayesian, *K*-means, outlier detection	Graduate student's information period from 1993 to 2007	Clustering student performance
31	[58]	To propose a hybrid model for intrusion detection to overcome difficulties with class dominance, force assignment and class problem.	*K*-means clustering	KDD cup 1999 data set	Clustering intrusion detection
32	[59]	To determine the different behavior patterns that are adopted by students in online discussion forums.	Agglomerative hierarchical clustering algorithm.	Students' activity in time series form	Clustering learner behaviour
33	[60]	Using DM algorithms into building mirroring tools to help small long-term teams improve their group work skills	*K*-means & EM algorithms in Weka	The data consisted of TRAC [61] Usage traces for 43 students working in seven groups with approximately 15,000 events	Clustering in CSCL
34	[62]	To discover and capture effective or ineffective student behaviors while interacting with the system	*K*-means clustering	Learner logged and eye-tracking data	Clustering Student modeling
35	[63]	To identify how reflective dialogues, predict student problem solving abilities	Hierarchical clustering and *K*-means clustering	Andes Physics dataset from PSLC Data shop	Clustering classroom learning

It is noteworthy to mention that clustering approach has been applied to different variables within the context of education. In the following sections, we make an attempt to present all these different educational variables to which clustering has been applied with successful results. The total research paper count is 166. The papers cited in table II, III, IV, V and VI are from five databases, namely, IEEEXplore, ACM Digital, JEDM, ProQuest Educational Journal and Science Direct. The search criteria are shown in section IV. Also as shown in table III, it is interesting to note that the maximum number of papers (more than 10) have been published in categories (EDM, e-learning and learning styles), while fewer than five papers have been published in categories (Annotation, classroom decoration, concept clustering, education affordability, exam failure, exam scheduling, Intelligent Tutoring System's (ITS), self-organizing map, semantic web in education, student motivation, student profiling and classroom learning). These areas provide the scope for improvement as well as areas for future research.

We have clustered various research works that have been conducted exclusively within educational attributes related to clustering algorithms and is shown in Table III.

TABLE III
EDUCATIONAL DATA CLUSTERING RESEARCH PAPERS AND THEIR THRUST AREAS

Educational data type & Clustering	References
Annotation	[64]
Classroom decoration	[65], [66], [49]
Collaborative learning	[42], [67], [43], [60], [68]
Data clustering – Review & Survey	[69], [70], [71], [72], [73], [74], [75], [76], [77]
Education affordability	[78]
Educational Data Mining (EDM)- Review & Survey	[79], [6], [1], [11]
Educational Data Mining (EDM) & Clustering	[79], [47], [80], [81], [82], [46], [83], [11], [84], [85], [86], [87], [50], [56], [88], [89], [90], [26], [91], [68], [56], [92], [93], [48], [94], [95], [96], [97], [98], [99]
E-Learning & Clustering	[100], [47], [77], [101], [30], [31], [32], [44], [33], [102], [103], [104], [105], [84], [106], [107], [34], [35], [36], [37], [38], [91], [39], [108], [109], [110], [41]
Examination failure & Clustering	[4], [48], [109]
Examination scheduling/Timetabling	[111], [112]
Intelligent Tutor System & Clustering	[113], [114]
Learning portfolio & Clustering	[41], [115], [116], [117]
Learning style & Clustering	[118], [80], [119], [120], [121], [122], [123], [124], [125], [126], [127], [128], [129]
Self-Organizing Map (SOM) & Clustering	[130], [131], [132]
Semantic web in education & Clustering	[44], [58]
Student modeling & Clustering	[18], [130], [133], [21], [101], [85], [134]
Student motivation & Clustering	[135]
Student profiling & Clustering	[55], [56], [136], [137], [138], [139]

We will now provide a detailed analysis on various aspects of educational attribute collated with the application of clustering algorithms to help improve the education system.

A. Analyzing Student Motivation, Attitude and Behavior

More often, students weak in mathematics would dread the mere notion of being asked by the teacher to sit in the front seat. Some common adages suggest that the back-benchers in a classroom are typically slow learners as compared to those who sit in the front seats. Students' seat selection in a classroom or lab environment and its implications on assessment was measured by Ivancevic, Celikovic & Lukovic [50]. *K*-means clustering was applied to an electronic log of 4096 records featuring information on student login/logout actions according to the time table of class meetings. After clustering, it was found that students with high levels of spatial deployment (seat selection) have 10% higher assessment scores as compared to students with low spatial choice.

Students typically write in the margins of books about their understanding of the text presented. This activity is called as 'annotation'. In one of a kind study proposed by Ying, *et al.* [64] two simple biology inspired approaches of chromosome behavior was applied to 40 students' annotations text. Then, they clustered the data based on the similarity between annotations using *K*-means clustering and hierarchical clustering methods. They found that their proposed approaches are more efficient than the generic hierarchical clustering algorithms.

Buehl & Alexander [135] studied students' epistemological beliefs about knowledge acquisition and their learning process. The objective of this research was to examine epistemological beliefs and students' achievement motivation. The unique aspect of this study is that rather than examining whether or not individual beliefs are related or co-related to performance and motivation; the authors tested different configurations of beliefs that were related to students' competence beliefs, achievement values and text-based learning. The sample size was 482 undergraduate students whose beliefs on knowledge, competency levels, and achievement values in history and mathematics were analyzed. Ward's minimum variance hierarchical clustering technique was used to analyze the data. The results revealed that students with different epistemological beliefs vary with their competency beliefs and achievement values. They suggested that future research may apply cluster analysis to different configurations of beliefs related to various aspects of student learning.

In a similar study a survey was conducted using Self Deterministic Theory (SDT) to measure student motivation towards learning and achievement by Dillon & Stolk [140]. The survey participants were 404 engineering students. The participant group consisted of 93 students in project-based material science course of an engineering college, 137 students in lecture-lab materials science course of a liberal arts university, and 174 students in lecture and lecture-lab course from a public University. Their data set comprised of 1278 complete survey responses and MCLUST method was applied to the data set. The results revealed situation-based motivation among engineering students but this motivation type could not be classified under the traditional intrinsic/extrinsic categories. Exactly like behavioural scientist who study their subjects' behaviour in order to understand them better, in a similar analogy the EDM scientist measure the behaviour of learners to design effective improvement solutions. Diverse studies have not been undertaken in domains such as student spatial

deployment, their motivation towards learning achievement, their epistemological belief about knowledge acquisition and inclination towards annotation behaviour. Yet, the aforementioned studies indicate that there are other similar areas that need to be explored and mined for the benefit of leaners, educators, and policy makers.

B. Understanding Learning Style

In 1971, David Kolb presented his infamous learning style theory called as "Experiential Learning Theory (ELT)" [141]. The term 'Experiential' means drawing knowledge based on previous experiences. In the same year, he also presented his Learning Style Inventory (LSI), a model used to assess differences in how individuals learn. Since then there have been various types of learning style inventories and learning theories. Some notable contributions are John Dewey's model of learning, as well as Piaget's model of learning and cognitive development. These learning style theories not only helped educators and researchers of the yesteryears but they continued to exert influence up to the present time.

Many studies reported the usage of learning styles in teaching to improve education quality Felder & Spurlin [142], Hawk & Shah [143]. Nowadays, learning style theories are used in an educational environment to enhance learning abilities of learners as well as teaching skills of educators. Looking at Table III, we notice that most publications are in e-learning. This indicates that considerable research work has been carried out in this field. It is obvious because the stage was already set, that is to say, the e-learning environment for the end-user was ready, the infrastructure in the form of internet was already in place and the database that held user activity was replete with data waiting to be mined by data scientists. However, little if any, research has been carried out on understanding learning styles of a learner in a spatial (classroom) environment using data mining methods such as clustering. 'Can easy accessibility to course material improve student learning or foster teaching in an e-learning environment?' is an interesting research question. In the following, we present notable research works that have contributed to answer this question.

A survey conducted at Warsaw School of Economics, where every semester more than 2000 students attend online lectures, showed that there are no significant improvements in student grades as compared to traditional classroom environment by Zajac [144]. This ground-breaking study stimulates another pertinent question, 'What factor is responsible for directly affecting learning and teaching so as to make or mar a learner's performance? The answer is in personalization of the learning content and individual's learning preferences, a fundamental factor in teaching and pedagogy. Every individual has his own learning preference as suggested by Felder [145]. Measuring individual's learning preferences is easy but how do we measure the learning preferences of all students in a class or semester? Fortunately, there exist mechanisms tailored for this specific purpose, aptly, called as Learning Style Inventories (LSIs). There are various types of LSIs available and the most acclaimed is Kolb's LSI [141].

In this paper, we aim to highlight research works that have applied clustering in various aspects of learning, therefore, we will not provide detailed discourse on LSI and it makes more sense to discuss clustering or any other data mining method as applied to LSI to improve learning. In this study by Rashid, *et al*. [125], where they applied statistical methods to determine LS based on human brain signals. The primary purpose of this study was to classify the participants' learning styles (LSs). A unique aspect of this study was analyzing the LS of the learner with psychoanalysis test using Mind Peak's Wave Rider instrument and brain signal processing. The effects of cognitive style on student learning in a Web Based Instruction (WBI) program using decision tree and *K*-means clustering method was studied by Chen, Chen & Liu [41] to automatically create student groups in a Computer Supported Collaborative Learning (CSCL) by considering individual learning styles as studied by Costaguta & de los Angeles Menini [146].

Much has been discussed so far regarding LS in critical reference to many of its attributes. But one imperative question remained unanswered and that is, 'How do you identify learning style of an individual?' This is best answered by Ahmad & Tasir [147]. As can be seen, most of the research works have focused on e-learning because of easy accessibility to data. This is in spite of the fact that there are several areas within learning styles as outlined above such as personalization of learning, learning style identification, and application of LSI in teaching that require further research, especially, in relation to data mining.

C. E-Learning

Perhaps the most notable research in the context of EDM has been done in reference to e-Learning. One of the reasons is the easy availability of data to analyse and infer from. In their paper Pardos, *et al*. [91] used a two-step analysis approach based on agglomerative hierarchical clustering to identify different participation profiles of learners in an online environment. Different levels of learner participation were measured by the number of posts, replies to the posts, frequency of threads posted, depth of the threads posted etc. Agglomerative Hierarchical clustering was used for this purpose. Data sets were adapted from online discussion forums of three different subjects in a virtual Telecommunications Degree (Electronic Circuits, Linear Systems Theory and Mathematics) over the period of three semesters (from February 2009 to July 2010). Thus, the whole data set involved a total amount of 672 learners distributed in eighteen different virtual classrooms and a total amount of 3842 posts. Total withdrawal and passing rates were 36.31% and 52.23%, respectively.

In another study conducted by Eranki & Moudgalya [37], 82 students from three engineering colleges were observed and *K*-means clustering was applied to their e-Learning data to investigate the influence of human characteristics on users' preferences while using WBeI. The sample size was 82 (51 male & 31 females). Then, Systematic Usability Scale (SUS) questionnaire was administered to the participants so that their perceptions on the use of Spoken tutorial interface could be identified. The SUS questionnaire was a 20.5 point Likert scaled questionnaire that was adapted to predict cognitive and affective data. By applying *K*-means clustering, the authors were able to find that expert computer users favoured multi-page, dynamic buttons and drop-down menus while the novice

users preferred single page, dynamic buttons and drop down menus. In Table IV, we show the research papers that have been published in clustering in e-learning sorted by the type of clustering algorithm used.

TABLE IV
RESEARCH PAPERS PUBLISHED IN CLUSTERING IN E-LEARNING

E-learning & Clustering		Published Papers
Non-Hierarchical Algorithm	*K*-means	[34], [37], [38], [30], [41], [106], [110], [32]
	C-means	[38]
	Fuzzy *K*-means	[73], [38]
	K-prototypes	[148]
	Fuzzy Clustering	[33], [32]
Hierarchical type algorithm	Agglomerative Clustering	[36], [31],[35]
	Markov Clustering	[39]
	Discrete Markov Model (DMM)	[30]

Research conducted by [104], discusses problem-solving behavior, different types of behavioral patterns of learners, and how these patterns can be automatically discovered. The purpose of applying this approach was to detect patterns based on targeted and automated clustering of users' problem-solving sequences as represented by Discrete Markov Models (DMMs). Data was taken from Andes Physics course of the USNA (2007-2009) from PSLC Data shop. The novelty of this research is that clustering has been applied at three different behavioral patterns. Level 1 (pattern driven), uses established predefined problem-solving styles and aims at discovering these patterns in student behavior. After clustering was performed on the data set of 8 clusters, two clusters with Trial and Error problem-solving style were identified. At level II (dimension-driven), the system tries to identify the given dimension and then helps in discovering the concrete styles along with these dimensions. Level III (open discovery), aims at the automatic discovery of both learning and dimensional style. A fundamental importance of this work is the employment of a set of optimization metrics that are applied on the achieved clusters to determine if the optimum cluster setting has been reached.

D. Collaborative Learning

Research on collaborative learning in an e-learning environment with students with mild disabilities was conducted by Chu, *et al*. [148]. It initially began with a focus on individuals in a group, later the focus shifted on the group itself and as the study progressed it was found that comparing the collaborative work with individual learning was more effective amongst the group participants. In a situation where a categorical variable has multi values the *K*-prototypes model as proposed by Huang [149] cannot be used. Therefore, one of the unique contributions of this study was that it proposed an enhanced clustering algorithm that used the *K*-prototypes model to cluster data with numerical, categorical single-values and categorical multi-values. Based on this clustering algorithm the researchers created context & content maps for creating their case-based reasoning recommendation system with semantic capabilities. This adaptive reasoning model enhanced teacher's practical knowledge and helped them to solve the student's learning problems. Collaboration is "the mutual engagement of participants' in a coordinated effort to solve a problem together" as suggested by Dillenbourg, *et al*. [150]. There have been various research works that have studied different variables pertaining to collaboration such as group size, composition of the group, communication channels within the group, interaction between peers and reward system in group work [150,151,152,153]. It has been argued that in order to understand and estimate the collaboration process, understanding the concept of collaborative learning is crucial by Mühlenbrock & Hoppe [154]. In this section, we will present and discuss research works that have implemented clustering algorithms to determine collaborative learning.

It is a learning method that requires learners to work together in groups or teams to reach a predetermined goal. It develops the abilities for interaction, fosters team building, enables sharing and cooperation and focusses on the collective perspective towards problem solving skills amongst learners within the group. As discussed in section 4.2, every student has a unique learning style. Chang, *et al*. [155] used Item Response Theory (IRT) to determine the students' ability and applied *K*-means clustering to group students together. This helped the teachers to adapt learning materials and teaching programs according to the student ability and aptitude. The experimental results showed that the average learning ability in the experimental group improved from 3.84 to 5.97. On the contrary, the learning ability in the control group only improved from 2.16 to 2.4.

One of the problems associated with this type of learning is that learners are not able to receive the appropriate support level from their collaborators. Anaya & Boticario [43] in their study found that clustering algorithms when applied to such data, build clusters according to learners' collaboration. So, active collaborators learn more in e-learning environment. Earlier related works have tackled this issue in different ways. Some researchers have looked at student collaboration from the perspective of experts evaluating collaborators' learning. Most other researchers approached it from collaborative information perspective. This information is then given to the learner/educator for their use. The novelty of this research is that by applying data mining methods, the researchers were able to recognize active and passive collaborators while learners were interacting with each other. The researchers used Expectation-Maximization (EM) clustering algorithm and Weka as their first step to build a data set. In this step, a data

set was built by applying statistical indicators to learner interaction within an online forum and was labeled accordingly. Over 100 students' data was derived from UNED European universities' largest online course using the dotLRN [156] platform. The number of students who took part in their research was 260 in 2006/2007 and 239 in 2007/2008. Examples of statistical indicators of learners were the number of threads posted or started by the learners, the average or the square variance etc. They were able to prove that highly active collaborators benefit more and their activities induce others too.

Learning in groups promotes learning motivation which increases student participation in learning activities and fosters good learning performance. In most cases, teachers would typically group students according to their grades. As such, students with poor grades may feel left-out. In an investigative study conducted by Perera, *et al*. [60], the objective was to improve teaching group work skills, facilitate effective team work by small groups, and work on substantial projects over several weeks by exploiting the electronic traces of group activity. For this purpose, *K*-means clustering was used along with WEKA and Euclidean distance measure. The data size of 43 students working in seven groups from TRAC [61] was 1.6MB in MYSQL format containing approximately 15,000 events. Also, EM clustering algorithm was used from Weka. Their cluster size for both *K*-means and EM was 3 clusters with 11 attributes and they obtained the same results, thus, proving that the choice of their attributes was good and without flaws because *K*-means is very sensitive to cluster sizes and also does not deal well with clusters with non-spherical shape and different sizes.

E. Educational Data Mining using Clustering

As we know that clustering algorithms can broadly be divided into hierarchical and non-hierarchical types. So, it would be easier if the research conducted could equally be partitioned according to the clustering algorithm used. This is shown in Table V following which we present a discussion on some of these works.

TABLE V
RESEARCH PAPERS PUBLISHED IN CLUSTERING IN EDUCATIONAL DATA

Educational data & Clustering		Published Papers
Non-Hierarchical Algorithm	*K*-means	[157], [158], [87], [90], [159]
	C-means	[47],
	Co-operative Particle Swarm Optimizer (PSO)	[94]
	Farthest First	[46]
	Expectation Maximization (EM)	[160]
Hierarchical Clustering Algorithm	Agglomerative Clustering	[59], [161]

Wook, *et al*. [46] have evaluated undergraduate students' academic performance on end of semester exam. They applied a combination of data mining methods such as Artificial Neural Network (ANN), Farthest-First method based on *K*-means clustering and Decision Tree as a classification approach. The data set comes from the faculty of science and defense technology, National Defense University of Malaysia (NUDM). Zheng and Jia, worked to improve the existing *K*-means clustering algorithm that has several drawbacks; In [157] they have stated that first, it is sensitive to the choice of the initial cluster centroids and may converge to the local optima; Second, the number of clusters needs to be determined in advance; and third, high dimensional data clustering takes a long time to finish. Co-operative Particle Swarm Optimizer (PSO) technique which is an improved version of *K*-means clustering is proposed by these researchers.

In an analytical study conducted by Parack, *et al*. [158] the applications of various DM techniques to student academic data has been provided. In this study, Apriori algorithm was applied to academic records of students to obtain the best association rules which help in student profiling. *K*-means clustering was used to group students categorically. The data is obtained from student academic record file; however, there is no mention of specific academic database being used. In this study by Zhiming & Xiaoli [81] worked on to identify the significant variables that affect and influence the performance of undergraduate students. The C-Means clustering method was used. But there is no mention of the data set used in the study. In another analytical study a group of researchers Zheng, *et al*. [30] attempted to cluster high dimensional educational data in this study. When traditional *K*-means, clustering is applied there is a huge computation cost involved, Therefore, to eliminate it, a new model is proposed that uses the Co-operative Particle Swarm Optimizer (PSO) frame to *K*-means clustering to reduce computation cost caused by *K*-means. (PSO) technique which is an improved version of *K*-means clustering is proposed.

In Fig. 1, we show the educational data clustering process. The first stage is the data pre-processing stage in which the researcher must first understand the domain and complexity of the educational dataset collected thereafter should be able to identify the attributes that have garbage or missing values. By garbage values we refer to values that are not marked to be present for the attribute.

Let us take an example, consider a nominal attribute 'student_response' with allowed values like 'yes' or 'no'.

Now, if this attribute is coded with a value like 'NA' then it should be treated as a garbage value and must be removed.

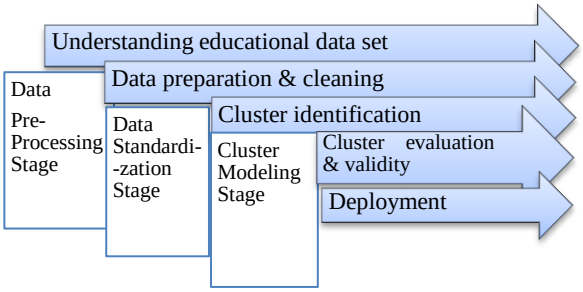

Fig. 1. Educational Data Clustering process

There should also be defined a standard to fill missing values. So, for example, referring to the above attribute 'student_response' the missing values could be filled as '?'. This activity is termed as data standardization. Once the data is cleaned, it should then be analyzed. Perhaps the easiest way is to determine relationships between various attributes that constitute the dataset. For example, Weka uses various machine learning algorithms (like Correlation attribute evaluator, One R attribute evaluator, gain ratio attribute evaluator, Principle component analysis attribute evaluator) that can easily determine the most significant attributes within the dataset. Once such significant attributes are found they can then be used to train and cluster the whole dataset to create data models. Post which new data having same or similar attributes can be applied to these data models to reveal interesting insights.

Chi, *et al*. [162] have conducted a study to determine student profiles based on their online browsing habits. The objectives of this research were two-fold. In the first step, they used content based filtering to extract keywords to obtain an article's characteristic descriptions. In the second step, Hierarchical *K*-means clustering was applied to this bag of keywords obtained in the first step. Web-pages were classified and then the researchers applied collaborative filtering to recommend web-pages. The research data consisted of viewing history of the web-pages over 30 days in ten computer labs. The number of pages viewed in 30 days was 42633 with 19 clusters.

Learning portfolios are records that are created during the learning process. Note taking, assignments, test paper reports, test papers etc. are examples of learning portfolio. In their analytical paper Chen, *et al*. [51] applied *K*-means, Farthest First and EM clustering algorithms and statistical *t*-test to the student portfolios of an e-learning system. Using clustering methods in this study they were able to cluster students' e-learning performance. Using t-test they were able to evaluate mid-term and final term exam performance of the clusters with high & low online learning frequency. 162 subjects used in this study were junior students of the department of computer engineering at Chung Yuan Christian University. This data was taken from i-learning [52] eLearning software being used in Taiwan. Their tests found that there was a positive correlation between students with high online eLearning frequency and higher scores. It was also found that the student portfolio of click times and duration of the study of learning materials at the beginning of the semester does not show any correlation with midterm and final term exam results. They also found that student participation in online discussion forums showed significant effect on their exam results.

In a similar analytical work conducted by Perera, *et al*. [60], *K*-means & EM clustering algorithms from WEKA was used to find group similarities. In this study their experiments revealed the same result for k=3 for *K*-means. Hierarchical agglomerative clustering with Euclidean distance was used for this purpose. The student teams were required to use TRAC [61] for online collaboration. TRAC is an open source, professional soft-ware development tracking system. The researchers collected the data over three semesters, for student cohorts in 2005 and 2006. The data size was 1.6 Mbytes in mySQL format and it contained approximately 15,000 events. The key contribution of this research is improved understandings of how to use data mining to build mirroring tools that can help small long-term teams improve their group work skills.

VI. DISCUSSION AND OPEN PROBLEM

So far, we see that subject specific research has been done but what about domain specific? For instance, how do institutions employ or apply data mining methods to improve institutional effectiveness? Zimmerman's educational model states that maintaining and monitoring students' academic record is an essential activity of an educational institutions. Anupama & Vijayalakshmi [86] applied classification and prediction algorithms, namely, decision table and One R algorithms on students' academic record from a previous semester to predict their performance in the current semester.

An educational institution maintains and stores various types of student data, it can range from student academic data to their personal records like parents' income, qualification and etc. In a research study by Tie, *et al*. [163] has proved that students performance can be predicted using a data set consisting of students' gender, parental education, their financial background. Chi, *et al*. [162] used Bayesian networks to predict student learning outcome based on attributes such as attendance, performance in class tests, assignments in this study. The researchers Knauf, *et al*. [165] have used the educational history of students for student modeling. While Dimokas, *et al*. [164] applied data mining methods like dimensional modeling into educational institutions be it a data warehousing solutions as applied in the department of Informatics of Aristotle University of Thessaloniki, storyboarding, or decision trees. While others like Nasiri, Minaei & Vafaei [166] used regression analysis and classification (CS5.0 algorithm which is a type of decision tree) to predict the academic dismissal of students and the GPA of graduated students in e-learning center. Considerable work has been done in e-learning. Perhaps the obvious reason is the easy availability of data. As this review indicates most of the e-learning software's are typically Moodle based. Also, if Table II is analyzed closely, it is noticed that there are certain areas like learner annotation, classroom decoration, learning outcome, exam failure, and examination scheduling/time-

tabling student motivation, student modeling and profiling that require more research work to be done in reference to application of clustering algorithms on them. One may argue that there is considerable literature on learning outcome or student modeling, no doubt there is but research work on examination failure and clustering is scarce and this caveat which is aptly shown in Table II requires to be filled. Organizing data into groups is a natural choice which we learn quite early in kindergarten. Similarly, organizing data into groups is predominant in many scientific fields. While numerous clustering algorithms have been published and new ones continue to proliferate; there has not been a single clustering algorithm till now that could dominate all others. In an education system, different users would interpret the same data differently for example, students, educators, school administrators, parents, and counsellors may hold various perspectives on examination report card data and each may be interested in generating different partitions or clusters from the same data set. Therefore, the viability of seeking a unified clustering algorithm would not be plausible. A clustering algorithm that satisfies the requirements of one user group may not satisfy the requirements of another user group. Given the inherent difficulty of understanding and applying clustering algorithm by a novice computer user, semi-supervised clustering techniques need to be developed in which the labeled data and paired constraints (user given) are applied to represent data and choose the appropriate function for educational data clustering. As shown in Table III little to almost negligible research has been conducted in areas such as learner annotation, effect of classroom decoration to augment learning and teaching, implications of education affordability, the inclusion of semantic web in education-its usability, learner motivation, timetabling, examination scheduling, student profiling and intelligent tutor systems. These are just a few of the many attributes that still require detailed research to be conducted from the computational perspective.

VII. CONCLUSION

This paper has presented over three decade's systematic review on clustering algorithm and its applicability and usability in the context of EDM. This paper has also outlined several future insights on educational data clustering based on the existing literatures reviewed, and further avenues for further research are identified. In summary, the key advantage of the application of clustering algorithm to data analysis is that it provides relatively an unambiguous schema of learning style of students given a number of variables like time spent on completing learning tasks, learning in groups, learner behavior in class, classroom decoration and student motivation towards learning. Clustering can provide pertinent insights to variables that are relevant in separating the clusters. Educational data is typically multi-level hierarchical and non-independent in nature, as suggested by Baker & Yacef [6] therefore a researcher must carefully choose the clustering algorithm that justifies the research question to obtain valid and reliable results.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, pp. 601-618, 2010.

[2] (2011, 01 July). *International Educational Data Mining Society*. Available: http://www.educationaldatamining.org/

[3] J. Ranjan and K. Malik, "Effective educational process: a data-mining approach," *Vine*, vol. 37, pp. 502-515, 2007.

[4] V. P. Bresfelean, M. Bresfelean, N. Ghisoiu, and C. A. Comes, "Determining students' academic failure profile founded on data mining methods," presented at the ITI 2008 - 30th International Conference on Information Technology Interfaces, 2008.

[5] J. Vandamme, -P.,Meskens, N., Superby, F.-, J, "Predicting Academic Performance by Data Mining Methods," *Education Economics*, vol. 15, pp. 405-419, 2007.

[6] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM-Journal of Educational Data Mining*, 2009.

[7] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," *Educause Review*, vol. 42, p. 40, 2007.

[8] J. Luan, "Data mining applications in higher education," *SPSS Executive*, vol. 7, 2004.

[9] S. H. Lin, "Data mining for student retention management," *Journal of Computing Sciences in Colleges*, vol. 27, pp. 92-99, 2012.

[10] T. Denley, "Austin Peay State University: Degree Compass," *EDUCAUSE Review Online.*, 2012.

[11] M. F. M. Mohsin, N. M. Norwawi, C. F. Hibadullah, and M. H. A. Wahab, "Mining the student programming performance using rough set," presented at the Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on, 2010.

[12] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert Systems with Applications*, vol. 33, pp. 135-146, 7/ 2007.

[13] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, pp. 1432-1462, 3/ 2014.

[14] O. R. Zaıane and J. Luo, "Web usage mining for a better web-based learning environment," in *Proceedings of conference on advanced technology for education*, 2001, pp. 60-64.

[15] O. R. Zaıane, "Building a recommender agent for e-learning systems," in *Computers in Education, 2002. Proceedings. International Conference on*, 2002, pp. 55-59.

[16] R. S. Baker, A. T. Corbett, and A. Z. Wagner, "Off-task behavior in the cognitive tutor classroom: when students game the system," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 383-390.

[17] P. Brusilovsky and C. Peylo, "Adaptive and intelligent web-based educational systems," *International Journal of Artificial Intelligence in Education*, vol. 13, pp. 159-172, 2003.

[18] J. E. Beck and B. P. Woolf, "High-level student modeling with machine learning," *Intelligent tutoring systems*, pp. 584-593, 2000.

[19] E. García, C. Romero, S. Ventura, and C. de Castro, "A collaborative educational association rule mining tool," *The Internet and Higher Education*, vol. 14, pp. 77-88, 2011.

[20] Y. H. Wang and H. C. Liao, "Data mining for adaptive learning in a TESL-based e-learning system," *Expert Systems with Applications*, vol. 38, pp. 6480-6485, 2011.

[21] M. E. Zorrilla, E. Menasalvas, D. Marin, E. Mora, and J. Segovia, "Web usage mining project for improving web-based learning sites," in *Computer Aided Systems Theory–EUROCAST 2005*, ed: Springer, 2005, pp. 205-210.

[22] T. S. Madhulatha, "An overview on clustering methods," *arXiv preprint arXiv:1205.1117*, 2012.

[23] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Englewood Cliffs,NJ, USA: Prentice-Hall, Inc., 1988.

[24] S. Sagiroglu and D. Sinanc, "Big Data: A Review," *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (Cts)*, pp. 42-47, 2013.

[25] J. Manyika, M. Chui, B. Brown, and J. Bughin, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[26] S. Parack, Z. Zahid, and F. Merchant, "Application of data mining in educational databases for predicting academic trends and patterns," in *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on*, 2012, pp. 1-4.

[27] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology*, vol. 51, pp. 7-15, 2009.

[28] H. M. Chen and M. D. Cooper, "Using clustering techniques to detect usage patterns in a Web-based information system," *Journal of the American Society for Information Science and Technology*, vol. 52, pp. 888-904, 2001.

[29] N. A. Rashid, M. N. Taib, S. Lias, N. Sulaiman, Z. H. Murat, and R. S. S. A. Kadir, "Learners' Learning Style Classification related to IQ and Stress based on EEG," *Procedia - Social and Behavioral Sciences*, vol. 29, pp. 1061-1070, / 2011.

[30] Q. Zheng, J. Ding, J. Du, and F. Tian, "Assessing Method for E-Learner Clustering," presented at the Computer Supported Cooperative Work in Design, 2007. CSCWD 2007. 11th International Conference on, 2007.

[31] P. Dradilova, J. Martinovic, K. Slaninova, and V. Snasel, "Analysis of Relations in eLearning," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, 2008, pp. 373-376.

[32] T. Feng, W. Shibin, Z. Cheng, and Z. Qinghua, "Research on e-learner personality grouping based on fuzzy clustering analysis," in *Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on*, Xi'an, 2008, pp. 1035-1040.

[33] C. Jili, H. Kebin, W. Feng, and W. Huixia, "E-learning Behavior Analysis Based on Fuzzy Clustering," in *Genetic and Evolutionary Computing, 2009. WGEC '09. 3rd International Conference on*, Guilin, 2009, pp. 863-866.

[34] S. B. Aher and L. Lobo, "Applicability of Data Mining Algorithms for Recommendation System in E-Learning," presented at the ICACCI '12: Proceedings of the International Conference on Advances in Computing, Communications and Informatics, 2012.

[35] P. D. Antonenko, S. Toy, and D. S. Niederhauser, "Using cluster analysis for data mining in educational technology research," *ETR&D-Educational Technology Research and Development*, vol. 60, pp. 383-398, 2012.

[36] G. Cobo, D. García-Solórzano, J. A. Morán, E. Santamaría, C. Monzo, and J. Melenchón, "Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums," in *2nd International Conference on Learning Analytics and Knowledge*, 2012.

[37] K. L. N. Eranki and K. M. Moudgalya, "Evaluation of Web Based Behavioral Interventions using Spoken Tutorials," presented at the IEEE Fourth International Conference on Technology for Education, 2012.

[38] F. Ghorbani and G. A. Montazer, "Learners grouping improvement in e-learning environment using fuzzy inspired PSO method," presented at the E-Learning and E-Teaching (ICELET), 2012 Third International Conference on, 2012.

[39] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A Clustering Methodology of Web Log Data for Learning Management Systems," *Educational Technology & Society*, vol. 15, pp. 154-167, 2012.

[40] C. Romero, M. I. López, J. M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458-472, 2013.

[41] C. M. Chen, Y. Y. Chen, and C. Y. Liu, "Learning Performance Assessment Approach Using Web-Based Learning Portfolios for E-learning Systems," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, pp. 1349-1359, 2007.

[42] C. Manikandan, A. S. Meenakshi Sundaram, and M. Mahesh Babu, "Collaborative E-Learning for Remote Education; An Approach For Realizing Pervasive Learning Environments," presented at the Information and Automation, 2006. ICIA 2006. International Conference on, 2006.

[43] A. R. Anaya and J. G. Boticario, "Clustering Learners according to their Collaboration," presented at the Computer Supported Cooperative Work in Design, 2009. CSCWD 2009. 13th International Conference on, 2009.

[44] C. T. Huang, W. T. Lin, S. T. Wang, and W. S. Wang, "Planning of educational training courses by data mining: Using China Motor Corporation as an example," *Expert Systems with Applications*, vol. 36, pp. 7199-7209, 2009.

[45] W. C. Chang, T. H. Wang, and M. F. Li, "Learning Ability Clustering in Collaborative Learning," *Journal of Software*, vol. 5, pp. 1363-1370, 2010.

[46] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques," *Second International Conference on Computer and Electrical Engineering, Vol 2, Proceedings*, pp. 357-361, 2009.

[47] A. Salazar, J. Gosalbez, I. Bosch, R. Miralles, and L. Vergara, "A case study of knowledge discovery on academic achievement, student desertion and student retention," presented at the Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on, 2004.

[48] A. Dharmarajan and T. Velmurugan, "Applications of Partition based Clustering Algorithms: A Survey," presented at the Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on, 2013.

[49] M. V. Almeda, P. Scupelli, R. S. Baker, M. Weber, and A. Fisher, "Clustering of design decisions in classroom visual displays," in *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge*, 2014, pp. 44-48.

[50] V. Ivancevic, M. Celikovic, and I. Lukovic, "The Individual Stability of Student Spatial Deployment and its Implications," presented at the Computers in Education (SIIE), 2012 International Symposium on, 2012.

[51] C. M. Chen, C. Y. Li, T. Y. Chan, B. S. Jong, and T. W. Lin, "Diagnosis of students' online learning portfolios," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*, 2007, pp. T3D-17-T3D-22.

[52] (2014). *e-learn*. Available: http://demo.learn.com.tw

[53] C. Li and J. Yoo, "Modeling Student Online Learning Using Clustering," in *Proceedings of the 44th annual Southeast regional conference*, 2006.

[54] R. S. Baker and S. M. Gowda, "Towards automatically detecting whether student learning is shallow," *Intelligent Tutoring* , 2012.

[55] C. C. Chi, C. H. Kuo, M. Y. Lu, and N. L. Tsao, "Concept-Based Pages Recommendation by Using Cluster Algorithm," presented at the Advanced Learning Technologies, 2008. ICALT '08. Eighth IEEE International Conference on, 2008.

[56] E. Trandafili, A. Allkoci, E. Kajo, and A. Xhuvani, "Discovery and evaluation of student's profiles with machine learning," *Proceedings of the Fifth Balkan Conference in Informatics on - BCI '12*, pp. 174-174, 2012.

[57] M. M. T. Tair and A. M. El-Halees, "Mining educational data to improve students' performance: A case study," *International Journal of Information and Communication Technology Research*, vol. 2, 2012.

[58] K. K. Bharti, S. Shukla, and S. Jain, "Intrusion detection using clustering," *Proceeding of the Association of Counseling Center Training Agencies (ACCTA)*, vol. 1, 2010.

[59] G. Cobo, D. García-Solórzano, E. Santamaria, J. A. Morán, J. Melenchón, and C. Monzo, "Modeling Students' Activity in Online Discussion Forums: A Strategy based on Time Series and Agglomerative Hierarchical Clustering," in *EDM*, 2011, pp. 253-258.

[60] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaiane, "Clustering and Sequential Pattern Mining of Online Collaborative Learning Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 759-772, June 2009.

[61] (2014, 10 June). *Welcome to the Trac Open Source Project*. Available: http://trac.edgewall.org

[62] S. Amershi and C. Conati, "Combining unsupervised and supervised classification to build user models for exploratory learning environments," *Journal of Educational Data Mining and Knowledge Discovery*, 2009.

[63] S. A. Sardareh, S. Aghabozorgi, and A. Dutt, "Reflective Dialogues and Students' Problem Solving Ability Analysis Using Clustering," presented at the The 3rd International Conference on Computer Engineering and Mathematical Sciences (ICCEMS 2014), Langkawi, Malaysia, 2014.

[64] K. Ying, M. Chang, A. F. Chiarella, and J. S. Heh, "Clustering Students based on Their Annotations of a Digital Text," *2012 Ieee Fourth International Conference on Technology for Education (T4e)*, pp. 20-25, 2012.

[65] J. C. Turner, P. K. Thorpe, and D. K. Meyer, "Students' reports of motivation and negative affect: A theoretical and empirical analysis," *Journal of Educational Psychology*, vol. 90, pp. 758-771, Dec 1998.

[66] R. Martinez-Maldonado, K. Yacef, and J. Kay, "Data Mining in the Classroom: Discovering Groups' Strategies at a Multi-tabletop Environment," *Proc. EDM*, vol. 2013, 2013.

[67] H. Mahdi and S. S. Attia, "Finding Candidate Helpers in Collaborative E-Learning using Rough Sets," presented at the Computer Information Systems and Industrial Management Applications, 2008. CISIM '08. 7th, 2008.

[68] G. Siemens and R. S. Baker, "Learning analytics and educational data mining: towards communication and collaboration," presented at the LAK '12: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 2012.

[69] O. A. Abbas, "Comparison between Data Clustering Algorithms," *The International Arab Journal of Information Technology*, vol. 5, pp. 320-325, 2008.

[70] A. P. de Barros Borges Reis Figueira, "A Repository with Semantic Organization for Educational Content," presented at the Eighth IEEE International Conference on Advanced Learning Technologies, 2008.

[71] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering algorithms and validity measures," *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management. SSDBM 2001*, pp. 3-22, 2001.

[72] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping multidimensional data*, ed: Springer, 2006, pp. 25-71.

[73] J. W. Zhao, S. M. Gu, and L. He, "A Novel Approach to Clustering Access Patterns in E-learning Environment," presented at the Education Technology and Computer (ICETC), 2010 2nd International Conference on, 2010.

[74] M. A. Dala and N. D. Harale, "A Survey on Clustering In Data Mining," presented at the International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET, Mumbai, India, 2011.

[75] F. Xhafa, J. J. Ruiz, S. Caballe, E. Spaho, L. Barolli, and R. Miho, "Massive Processing of Activity Logs of a Virtual Campus," presented at the Third International Conference on Emerging Intelligent Data and Web Technologies, 2012.

[76] A. Nagpal, A. Jatain, and D. Gaur, "Review based on data clustering algorithms," *Information & Communication*, pp. 298-303, 2013.

[77] R. Xu and D. Wunsch, "Survey of clustering algorithms," *Neural Networks, IEEE Transactions on*, vol. 16, pp. 645 - 678, May 2005 2005.

[78] A. Arulselvan, P. Mendoza, V. Boginski, and P. M. Pardalos, "Predicting the Nexus between Post-Secondary Education Affordability and Student Success: An Application of Network-based Approaches," presented at the Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in, Athens, 2009.

[79] F. R. Lin, C. H. Chen, and K. L. Tsai, "Discovering Group Interaction Patterns in a Teachers Professional Community," presented at the System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on, 2003.

[80] C. Romero, S. Ventura, and E. Garcia, "Data mining in course management systems: Moodle case study and tutorial," *Computers & Education*, vol. 51, pp. 368-384, Aug 2008.

[81] Q. Zhiming and W. Xiaoli, "Application of RS and Clustering Algorithm in Distance Education," presented at the 2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing, 2008.

[82] F. Siraj and M. A. Abdoulha, "Uncovering Hidden Information Within University's Student Enrollment Data Using Data Mining," *2009 Third Asia International Conference on Modelling & Simulation, Vols 1 and 2*, pp. 413-418, 2009.

[83] S. P. Kumar and K. S. Ramaswami, "Fuzzy *K*- Means Cluster Validation for Institutional Quality Assessment," presented at the Communication and Computational Intelligence (INCOCCI), 2010 International Conference on, 2010.

[84] M. Zorrilla, D. García, and E. Alvarez, "A Decision Support System to improve e-Learning Environments," presented at the Proceedings of the 2010 EDBT/ICDT Workshops, 2010.

[85] M. C. Desmarais and R. S. Baker, "A review of recent advances in learner and skill modeling in intelligent learning environments," *User Modeling and User-Adapted Interaction*, vol. 22, pp. 9-38, 18 October 2011 2011.

[86] K. S. Anupama and M. N. Vijayalakshmi, "Mining of student academic evaluation records in higher education," *2012 International Conference on Recent Advances in Computing and Software Systems*, pp. 67-70, 2012.

[87] A. Banumathi and A. Pethalakshmi, "A Novel Approach for Upgrading Indian Education by Using Data Mining Techniques," *2012 IEEE International Conference on Technology Enhanced Education (ICTEE 2012)*, 2012.

[88] M. H. Abdous, W. He, and C. J. Yen, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade," *Educational Technology & Society*, vol. 15, pp. 77-88, 2012.

[89] W. Jin, T. Barnes, J. Stamper, M. J. Eagle, M. W. Johnson, and L. Lehmann, "Program Representation for Automatic Hint Generation for a Data-Driven Novice Programming Tutor," in *Intelligent Tutoring Systems*. vol. 7315, S. A. C. cerri@lirmm.fr, W. J. C. william.j.clancey@nasa.gov, G. P. papadour@cs.teicrete.gr, and K. Panourgia, Eds., ed, 2012, pp. 304-309.

[90] S. V. Lahane, M. U. Kharat, and P. S. Halgaonkar, "Divisive approach of Clustering for Educational Data," *Proceedings of the 2012 Fifth International Conference on Emerging Trends in Engineering and Technology (ICETET 2012)*, pp. 191-195, 2012.

[91] Z. A. Pardos, S. Trivedi, N. T. Heffernan, and G. N. Sárközy, "Clustered Knowledge Tracing," presented at the ITS'12: Proceedings of the 11th international conference on Intelligent Tutoring Systems, 2012.

[92] B. Azarnoush, J. M. Bekki, G. C. Runger, B. L. Bernstein, and R. K. Atkinson, "Toward a Framework for Learner Segmentation," *JEDM-Journal of Educational Data Mining*, vol. 5, 2013.

[93] S. Bryfczynski, R. P. Pargas, M. M. Cooper, M. Klymkowsky, and B. C. Dean, "Teaching Data Structures with beSocratic," presented at the ITiCSE '13: Proceedings of the 18th ACM conference on Innovation and technology in computer science education, 2013.

[94] K. Govindarajan, T. S. Somasundaram, and V. S. Kumar, "Continuous Clustering in Big Data Learning Analytics," *2013 IEEE Fifth International Conference on Technology for Education (t4e 2013)*, pp. 61-64, 2013.

[95] S. K. Mohamad and Z. Tasir, "Educational data mining: A review," *Procedia-Social and Behavioral Sciences*, vol. 97, pp. 320-324, 6 Novermber 2013 2013.

[96] C. Troussas, M. Virvou, J. Caro, and K. J. Espinosa, "Mining relationships among user clusters in Facebook for language learning," presented at the Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on, 2013.

[97] A. Bogarín, C. Romero, R. Cerezo, and M. Sánchez-Santillán, "Clustering for improving educational process mining," *Proceedins of the Fourth International Conference on Learning Analytics And Knowledge - LAK '14*, pp. 11-15, 2014.

[98] S. Aghabozorgi, H. Mahroeian, A. Dutt, T. Y. Wah, and T. Herawan, "An Approachable Analytical Study on Big Educational Data Mining," in *Computational Science and Its Applications–ICCSA 2014*, ed: Springer, 2014, pp. 721-737.

[99] B. Chakraborty, K. Chakma, and A. Mukherjee, "A density-based clustering algorithm and experiments on student dataset with noises using Rough set theory," in *Engineering and Technology (ICETECH), 2016 IEEE International Conference on*, 2016, pp. 431-436.

[100] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, pp. 354-359, 1983.

[101] K. Pata, M. Pedaste, and T. Sarapuu, "The formation of learners' semiosphere by authentic inquiry with an integrated learning object "Young Scientist"," *Computers & Education*, vol. 49, pp. 1357-1377, 12/ 2007.

[102] L. Zhuhadar, O. Nasraoui, R. Wyatt, and E. Romero, "Multi-model Ontology-based Hybrid Recommender System in E-learning Domain," presented at the Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09. IEEE/WIC/ACM International Joint Conferences on, 2009.

[103] Z. Jian-Wen, S. M. Gu, and L. He, "A Novel Approach to Clustering Access Patterns in E-learning Environment," *Education Technology and Computer (ICETC), 2010 2nd International Conference on*, vol. 1, pp. V1-393, 2010.

[104] M. Kock and A. Paramythis, "Towards Adaptive Learning Support on the Basis of Behavioural Patterns in Learning Activity Sequences," presented at the Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on, 2010.

[105] F. Xhafa, S. Caballe, L. Barolli, A. Molina, and R. Miho, "Using Bi-clustering Algorithm for Analyzing Online Users Activity in a Virtual Campus," presented at the Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on, 2010.

[106] T. Chellatamilan, M. Ravichandran, R. M. Suresh, and G. Kulanthaivel, "Effect of Mining educational Data to improve Adaptation of learning in e-Learning System," presented at the Sustainable Energy and Intelligent Systems (SEISCON 2011), International Conference on, 2011.

[107] M. Köck and A. Paramythis, "Activity sequence modelling and dynamic clustering for personalized e-learning," *User Modeling and User-Adapted Interaction*, vol. 21, pp. 51-97, 2011.

[108] K. Govindarajan, T. S. Somasundaram, and V. S. Kumar, "Particle Swarm Optimization (PSO)-based Clustering for Improving the Quality of Learning using Cloud Computing," *2013 Ieee 13th International Conference on Advanced Learning Technologies (ICALT 2013)*, pp. 495-497, 2013.

[109] H. Hani, H. Hooshmand, and S. Mirafzal, "Identifying the Factors Affecting the Success and Failure of E-learning Students Using Cluster Analysis," presented at the e-Commerce in Developing Countries: With Focus on e-Security (ECDC), 2013 7th Intenational Conference on, Kish Island, 2013.

[110] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses," presented at the LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge, 2013.

[111] S. Shatnawi, K. Al-Rababah, and B. Bani-Ismail, "Applying a Novel Clustering Technique Based on FP- Tree to University Timetabling Problem: A Case Study," presented at the Computer Engineering and Systems (ICCES), 2010 International Conference on, 2010.

[112] T. Van To and S. San Win, "Clustering approach to examination scheduling," presented at the Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference, 2010.

[113] F. Bouchet, J. M. Harley, G. J. Trevors, and R. Azevedo, "Clustering and Profiling Students According to their Interactions with an Intelligent Tutoring System Fostering Self-Regulated Learning," *JEDM-Journal of Educational Data Mining*, vol. 5, 2013.

[114] S. P. Kumar and K. S. Ramaswami, "Fuzzy *K*- Means Cluster Validation for Institutional Quality Assessment," presented at the Communication and Computational Intelligence (INCOCCI), 2010 International Conference on, 2010.

[115] W. Wang, J. F. Weng, J. M. Su, and S. S. Tseng, "Learning Portfolio Analysis and Mining for SCORM Compliant Environment," *Frontiers in Education (FIE)*, vol. 9, pp. T2C-17, 2004.

[116] C. M. Chen and Y. Y. Chen, "Learning Performance Assessment Approach Using Learning Portfolio for E-learning Systems," presented at the Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on, 2005.

[117] C. M. Chen, C. H. Ma, B. S. Jong, Y. T. Hsia, and T. W. Lin, "Using Data Mining to Discover the Correlation between Web Learning Portfolios and Achievements," presented at the Frontiers in Education Conference, 2008. FIE 2008. 38th Annual, 2008.

[118] S. Y. Cheng, C. S. Lin, H. H. Chen, and J. S. Heh, "Learning and diagnosis of individual and class conceptual perspectives: an intelligent systems approach using clustering techniques," *Computers & Education*, vol. 44, pp. 257-283, 4/ 2005.

[119] C. M. Chen and M. C. Chen, "Mobile formative assessment tool based on data mining techniques for supporting web-based learning," *Computers & Education*, vol. 52, pp. 256-273, 1/ 2009.

[120] C. Romero, S. Ventura, A. Zafra, and P. D. Bra, "Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems," *Computers & Education*, vol. 53, pp. 828-840, 11/ 2009.

[121] H. Guruler, A. Istanbullu, and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," *Computers & Education*, vol. 55, pp. 247-254, 8/ 2010.

[122] G. Forestier, P. Gançarski, and C. Wemmert, "Collaborative clustering with background knowledge," *Data & Knowledge Engineering*, vol. 69, pp. 211-228, 2, 2010.

[123] S. T. Levy and U. Wilensky, "Mining students' inquiry actions for understanding of complex systems," *Computers & Education*, vol. 56, pp. 556-573, 2011.

[124] S. Martin, G. Diaz, E. Sancristobal, R. Gil, M. Castro, and J. Peire, "New technology trends in education: Seven years of forecasts and convergence," *Computers & Education*, vol. 57, pp. 1893-1906, 11, 2011.

[125] N. A. Rashid, M. N. Taib, S. Lias, N. Bin Sulaiman, and Z. H. Murat, "EEG Theta and Alpha Asymmetry analysis of Neuroticism-bound Learning Style," presented at the 3rd International Congress on Engineering Education (ICEED), 2011.

[126] W. He, "Examining students' online interaction in a live video streaming environment using data mining and text mining," *Computers in Human Behavior*, vol. 29, pp. 90-102, 1/ 2013.

[127] C. F. Lin, Y. C. Yeh, Y. H. Hung, and R. I. Chang, "Data mining for providing a personalized learning path in creativity: An application of decision trees," *Computers & Education*, vol. 68, pp. 199-210, 10, 2013.

[128] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52-78, 3/ 2014.

[129] B. E. Vaessen, F. J. Prins, and J. Jeuring, "University students' achievement goals and help-seeking strategies in an intelligent tutoring system," *Computers & Education*, vol. 72, pp. 196-208, 3, 2014.

[130] C. S. Lee and Y. P. Singh, "Student modeling using principal component analysis of SOM clusters," *IEEE International Conference on Advanced Learning Technologies, Proceedings*, pp. 480-484, 2004.

[131] R. Saadatdoost, A. T. H. Sim, and H. Jafarkarimi, "Application Of Self Organizing Map For Knowledge Discovery Based In Higher Education Data," presented at the Research and Innovation in Information Systems (ICRIIS), 2011 International Conference on, Kuala Lumpur, 2011.

[132] A. S. Sabitha and D. Mehrotra, "User Centric Retrieval of Learning Objects in LMS," *2012 Third International Conference on Computer and Communication Technology (ICCCT)*, pp. 14-19, 2012.

[133] A. Merceron and K. Yacef, "Mining Student Data Captured from a Web-Based Tutoring Tool: Initial Exploration and Results," *Journal of Interactive Learning Research*, vol. 15, pp. 319-345, 2004 2010-06-08 2004.

[134] D. Zakrzewska, *Using clustering technique for students' grouping in intelligent e-learning systems*: Springer, 2008.

[135] M. M. Buehl and P. A. Alexander, "Motivation and performance differences in students domain-specific epistemological belief profiles," *American Educational Research Journal*, vol. 42, pp. 697-726, Win 2005.

[136] A. F. Wise, J. Speer, F. Marbouti, and Y.-T. Hsiao, "Broadening the notion of participation in online discussions: examining patterns in learners' online listening behaviors," *Instructional Science*, vol. 41, pp. 323-343, 2013.

[137] T. M. Khan, F. Clear, and S. S. Sajadi, "The relationship between educational performance and online access routines: analysis of students' access to an online discussion forum," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 2012, pp. 226-229.

[138] A. M. Bliuc, R. Ellis, P. Goodyear, and L. Piggott, "Learning through face-to-face and online discussions: Associations between students' conceptions, approaches and academic performance in political science," *British Journal of Educational Technology*, vol. 41, pp. 512-524, 2010.

[139] L. Talavera and E. Gaudioso, "Mining student data to characterize similar behavior groups in unstructured collaboration spaces," in *Workshop on artificial intelligence in CSCL. 16th European conference on artificial intelligence*, 2004, pp. 17-23.

[140] A. Dillon and J. Stolk, "The students are unstable! Cluster analysis of motivation and early implications for educational research and practice," *2012 Frontiers in Education Conference Proceedings*, pp. 1-6, 2012.

[141] D. A. Kolb, *Experiential learning: Experience as the source of learning and development* vol. 1, 1984.

[142] R. M. Felder and J. Spurlin, "Applications, reliability and validity of the index of learning styles," *International journal of engineering education*, vol. 21, pp. 103-112, 2005.

[143] T. F. Hawk and A. J. Shah, "Using learning style instruments to enhance student learning," *Decision Sciences Journal of Innovative Education*, vol. 5, pp. 1-19, 2007.

[144] M. Zajac, "Using learning styles to personalize online learning," *Campus-wide information systems*, vol. 26, pp. 256-265, 2009.

[145] R. M. Felder, "Matters of style," *ASEE prism*, vol. 6, pp. 18-23, 1996.

[146] R. Costaguta and M. de los Angeles Menini, "An assistant agent for group formation in CSCL based on student learning styles," presented at the Proceedings of the 7th Euro American Conference on Telematics and Information Systems, Valparaiso, Chile, 2014.

[147] N. Ahmad and Z. Tasir, "Threshold Value in Automatic Learning Style Detection," *Procedia - Social and Behavioral Sciences*, vol. 97, pp. 346-352, 11/6/ 2013.

[148] H. C. Chu, T. Y. Chen, C. J. Lin, M. J. Liao, and Y. M. Chen, "Development of an adaptive learning case recommendation approach for problem-based e-learning on mathematics teaching for students with mild disabilities," *Expert Systems with Applications*, vol. 36, pp. 5456-5468, 4/ 2009.
[149] Z. Huang, "Extensions to the *K*-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, pp. 283-304, 1998.
[150] P. Dillenbourg, M. J. Baker, A. Blaye, and C. O'Malley, "The evolution of research on collaborative learning," *Learning in Humans and Machine: Towards an interdisciplinary learning science.*, pp. 189-211, 1995.
[151] D. Adams and M. Hamm, *Cooperative Learning: Critical Thinking and Collaboration Across the Curriculum*: ERIC, 1996.
[152] R. E. Slavin, "Synthesis of Research of Cooperative Learning," *Educational leadership*, vol. 48, pp. 71-82, 1991.
[153] G. Underwood, M. McCaffrey, and J. Underwood, "Gender differences in a cooperative computer-based language task," *Educational Research*, vol. 32, pp. 44-49, 1990.
[154] M. Mühlenbrock and U. Hoppe, "Computer supported interaction analysis of group problem solving," in *Proceedings of the 1999 conference on Computer support for collaborative learning*, 1999, p. 50.
[155] W.-C. Chang, S.-L. Chen, M.-F. Li, and J.-Y. Chiu, "Integrating IRT to Clustering Student's Ability with K-Means," in *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, 2009, pp. 1045-1048.
[156] (2014, 10 June 2014). *.LRN*. Available: http://dotlrn.org/
[157] X. Zheng and Y. Jia, "A study on educational data clustering approach based on improved particle swarm optimizer," in *IT in Medicine and Education (ITME), 2011 International Symposium on*, 2011, pp. 442-445.
[158] S. Parack, Z. Zahid, and F. Merchant, "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns," *2012 Ieee International Conference on Technology Enhanced Education (Ictee 2012)*, pp. 1-4, 2012.
[159] T. Zhixin, J. Rong, Z. Hong, and W. Zhaoqing, "The Research on Teaching Method of Basics Course of Computer based on Cluster Analysis," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, 2010, pp. 2001-2004.
[160] A. Bovo, S. Sanchez, O. Heguy, and Y. Duthen, "Clustering moodle data as a tool for profiling students," in *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, 2013, pp. 121-126.
[161] F. Wijayanto, "Indonesia education quality: Does distance to the capital matter? (A clustering approach on elementary school intakes and outputs qualities)," in *Science and Technology (TICST), 2015 International Conference on*, 2015, pp. 318-322.
[162] C.-C. Chi, C.-H. Kuo, M.-Y. Lu, and N.-L. Tsao, "Concept-Based Pages Recommendation by Using Cluster Algorithm," presented at the Eighth IEEE International Conference on Advanced Learning Technologies, Santander, Cantabria, 2008.
[163] Z. Tie, R. Jin, H. Zhuang, and Z. Wang, "The Research on Teaching Method of Basics Course of Computer based on Cluster Analysis," *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, pp. 2001-2004, 2010.
[164] N. Dimokas, N. Mittas, A. Nanopoulos, and L. Angelis, "A Prototype System for Educational Data Warehousing and Mining," *Pci 2008: 12th Pan-Hellenic Conference on Informatics, Proceedings*, pp. 199-203, 2008.
[165] R. Knauf, Y. Sakurai, K. Takada, and S. Tsuruta, "A Case Study on Using Personalized Data Mining for University Curricula," *Proceedings 2012 IEEE International Conference on Systems, Man, and Cybernetics (Smc)*, pp. 3051-3056, 2012.
[166] M. Nasiri, B. Minaei, and F. Vafaei, "Predicting GPA and academic dismissal in LMS using educational data mining: A case mining," *E-Learning and E-Teaching (ICELET), 2012 Third International Conference on*, pp. 53-58, 14-15 Feb. 2012 2012.

Authors Biography

Ashish Dutt is a PhD Candidate at the Department of Information Systems of the Faculty of Computer Science and Information Technology at University of Malaya. He is a data analyst with over a decade experience in IT and education related domains. His research interest includes rough and soft set theory, DMKDD. He is an IEEE student member and acts as an active reviewer for many international journals.

Maizatul Akmar Ismail is a senior lecturer at the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya (UM), Malaysia. Her academic qualifications were obtained from University of Malaya for Bachelor and PhD degree, and University of Putra Malaysia for Masters. At present, she has more than fifteen years of teaching experience since she started her career in 2001 as a lecturer in the University of Malaya. Maizatul was involved in various research, leading to publication of a number of academic papers in the areas of Information Systems specifically on Semantic Web and Educational Technology. She has been actively publishing more than 40 conference papers in renowned local and international conferences. A number of her works were also published in reputable international journals. Maizatul has participated in many competitions and exhibitions to promote her research works. She actively supervises at all level of studies, from undergraduate to PhD. To date, she has successfully supervised three PhD and numbers of masters students to completion. She hopes to extend her research beyond Information Systems in her quest to elevate the the quality of teaching and learning.

Tutut Herawan is presently an associate professor at Department of Information Systems, University of Malaya. He received PhD degree in information technology in 2010 from Universiti Tun Hussein Onn Malaysia. He is also a principal researcher at AMCS Research Center, Indonesia. He has more than 12 years experiences as academic and also successfully supervised five PhD students. He presently supervises 15 Master and PhD students and has examined Master & PhD Theses. He is the executive editor of Malaysian Journal of Computer Science (ISI JCR with IF 0.405). He has also guest edited many special issues in many reputable international journals. He has edited five many books and published extensive articles in various book chapters, international journals and conference proceedings. He is an active reviewer for various journals. He delivered many keynote addresses, invited workshop and seminars and has been actively served as a chair, co-chair, program committee member and co-organizer for numerous international conferences/workshops. His research area includes soft computing, data mining, and information retrieval.