Educational Data Mining as a Trend of Data Mining in Educational System

Sachin R. Barahate
Department of Computer Engineering
Yadavrao Tasgaonkar College of Engineering and Management, Chandhai, Karjat.
University Mumbai, India.

ABSTRACT

Educational data mining is an emerging trend, concerned with developing methods for exploring the huge data that come from the educational system. This data is used to derive the knowledge which is useful in decision making. EDM methods are useful to measure the performance of students, assessment of students and study students' behavior etc. In recent years, Educational data mining has proven to be more successful at many of the educational statistics problems due to enormous computing power and data mining algorithms.

This paper surveys the history and applications of data mining techniques in the educational field. The objective is to introduce data mining to traditional educational system, web-based educational system, intelligent tutoring system, and e-learning. This paper describes how to apply the main data mining methods such as prediction, classification, relationship mining, clustering, and social area networking to educational data.

General Terms

Algorithms, Management, Measurement, Performance, Standardization, Theory.

Keywords

Educational Data Mining (EDM), Data Mining, Knowledge Discovery in Databases (KDD), Intelligent Tutoring System.

1. INTRODUCTION

The data mining has attracted a great deal of attention in the information technology industry, due to availability of large volume of data which is stored in various formats like files, texts, records, images, sounds, videos, scientific data and many new data formats. There is imminent need for turning such huge data into meaningful information and knowledge. The data collected from various applications require a proper data mining technique to extract the knowledge from large repositories for decision making. Data mining, also called Knowledge Discovery in Dabatabases (KDD), is the field of discovering novel and potentially useful information from large volume of data [45].

Data mining and knowledge discovery in databases are treated as synonyms, but data mining is actually a step in the process of knowledge discovery. The sequences of steps indentified in extracting knowledge from data are shown in Figure 1. The main functionality of data mining techniques is applying various methods and algorithms in order to discover and extract patterns of stored data. These interesting patterns are presented to the user and may be stored as new knowledge in knowledge base. Data mining and knowledge discovery applications have got a rich focus due to its significance in

decision making. Data mining has been used in areas such as database systems, data warehousing, statistics, machine learning, data visualization, and information retrieval. Data mining techniques have been introduced to new areas including neural networks, patterns recognition, spatial data analysis, image databases and many application fields such as business, economics and bioinformatics.
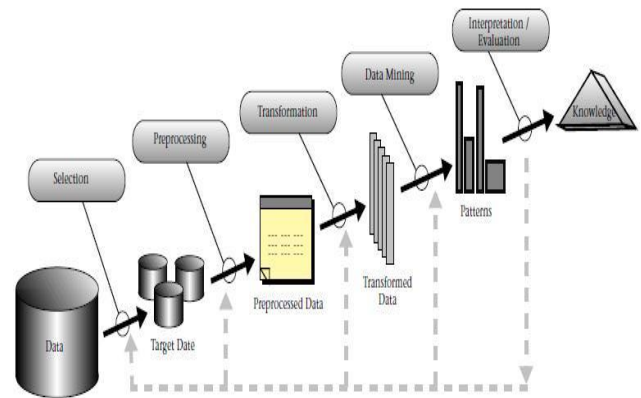

Figure 1. The steps of extracting knowledge from data.

The main objective of this paper is to survey the applications of data mining techniques/algorithms to traditional educational system. The first section is used to describe the history and current trends in the field of Educational Data Mining (EDM). I then review the different data mining techniques that have been applied in educational field grouping them by task. Finally, I discuss and summarize key applications of educational data mining.

2. EDUCATIONAL DATA MINING

The educational data mining community [48] defines educational data mining as, "Educational Data Mining (EDM) is 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 setting which they learn in".

There are increasing research interests in using data mining techniques in educational filed. This new emerging field, EDM, concerns with developing methods that discover knowledge from data originating from educational environments.

Educational data mining techniques often differ from traditional data mining techniques, in explicitly exploiting the multiple levels of meaningful hierarchy in educational data.

EDM focuses on collection, archiving, and analysis of data related to students' learning and assessment. The analysis performed in EDM research is often related to techniques drawn from variety of literatures [4], including psychometrics, machine learning, data mining, educational statistics, information visualization and computational modeling. The application of data mining in educational system is an interactive cycle of hypothesis formation, testing, and refinement as shown is Figure 2 [35]. Discovered knowledge should enter the loop of the system and guide, facilitate, and enhance learning as a whole. The system is used to turning data into knowledge as well as filtering mined knowledge for decision making.

3. EDM METHODS

Romero and Ventura, [36][37], and Baker categorize methods in educational data mining into the following general categories:

- Prediction
- Clustering
- Relationship mining
- Discovery with models
- Distillation of data for human judgment

Romero and Ventura's viewpoint is focused on applications of educational data mining to data. These methods are listed as web mining methods, and are quite prominent in mining web data and in mining other forms of educational data [12]. The first three categories of educational data mining method are largely acknowledged to be universal across types of data mining. The fourth and fifth categories achieve particular prominence within educational data mining.
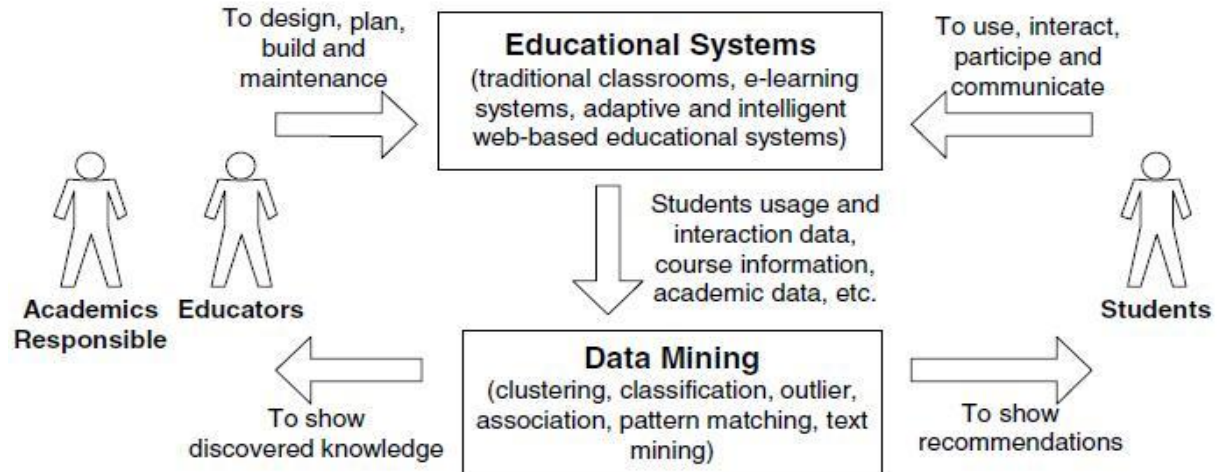
3.1 Prediction

The goal of prediction is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of data (predictor variables). Prediction has two key uses within educational data mining. In first type, prediction methods can be used to study what features of a model are important for prediction, giving information about underlying construct. This approach is basically used to analyze the students' performance [9].

In second type of usage, the prediction methods are used in order to predict what the output value would be in contexts where it is not desirable to directly obtain a label for that construct. Prediction can be classified in three types viz. Classification, Regression and Density estimation.

3.1.1 Classification

A classifier is mapping from (discrete or continuous) feature space X to discrete set of labels Y [16]. Classification or discriminant analysis is used to predict class labels (describes future situation). Classification is supervised technique which is used to label newly encountered (still unlabeled) patterns from a collection of labeled (pre-classified) patterns. Some popular classification methods include logistic regression, support vector machines and decision trees. The decision tree algorithm produces a tree-like structure of the model it produces. Form the tree it is then easy to generate rules in the form IF condition THEN outcome. It is basically a predictive model in which an instance is classified by following the path of satisfied condition from root until reaching a leaf, which will correspond to class label. Some of the most well-known decision tree algorithms are C4.5 and ID3.

Figure 2. The cycle of applying data mining in educational system.

3.1.2 Regression

Regression technique can also be adopted for prediction. In regression, the predicted variable is a continuous variable. Some popular regression methods within educational data mining include linear regression, neural networks and support vector machine regression. Many real-world educational data mining problems are not simply prediction. Therefore more complex techniques may be necessary to forecast future values using combination of the techniques (e.g. logistic regression, decision trees or neural networks). For example, the CART (Classification and Regression Tree) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees

(to forecast continuous response variables). Neural network too can create both classification and regression models.

3.1.3 Density estimation

In density estimation, the predicted variable is a probability density function. Density estimator can be based on variety of kernel functions, including Gaussian function. For each of the prediction methods, the input may be continuous or categorical. The particular predictive method is effective depending on the type of input variable used.

3.2 Clustering

Clustering is process of grouping objects into classes of similar objects [21]. It is an unsupervised classification or partitioning of patterns (observations, data items or future vectors) into groups or subsets (clusters) based on their locality and connectivity within N-dimensional space. Clustering is particularly useful in cases where the most common categories within the data set are not known in advance. In educational data mining, clustering has been used to group students according to their behavior [3]. The schools could be clustered together (to investigate similarities & differences between schools), students could be clustered together (to investigate similarities and differences between students). Some of the well known clustering methods are k-mean and expectation-maximization algorithm (EM-clustering) [14].

3.3 Relationship Mining

In relationship mining, the goal is to discover relationships between variables, in a data set with large number of variables. This may take the form of attempting to find out which variables are most strongly related/ associated with a single variable of particular interest. Broadly relationship mining is classified into four types: association rule mining, correlation mining, sequential pattern mining, and casual data mining.

3.3.1 Association rule mining

Association rule mining discovers relationships among attributes in data set, producing if-then statements concerning attribute-values [2]. Association rule mining is one of the important technique which aims at extracting, interesting correlations, frequent patterns, associations or casual structures among set of items in the transaction databases. An X => Y association rule expresses a close relationship between items (attribute-value) in a database with values of support and confidence. Association rule mining has been applied to educational data mining for: finding students' mistakes often occurring together while solving exercises [26] , finding out relationships in learners' behavior patterns [49], diagnosing students' learning problems and offers students advice [20] etc.

3.3.2 Correlation mining

In correlation mining, the goal is to find (positive or negative) linear correlations between variables. Correlation analysis is used to find the most strongly correlation attributes.

3.3.3 Sequential pattern mining

Sequential pattern mining is a more restrictive form of association rule mining in which the accessed item's order is taken into consideration. Sequential pattern mining [1] attempts to find inter-session patterns such as the presence of a set of items followed by another items in a time-ordered set of sessions or episodes. Sequential pattern mining has been applied to web-based educational systems. Sequential pattern can reveal which content has motivated the access to another contents, or how tools and contents are entwined in the learning process. Zaine and Luo [50] propose the discovery of useful patterns based on restrictions, to help educators evaluates students' activities in web courses. Paul and Donnelan [30] discover and compare with expected behavioral patterns specified by the instructor that describe an ideal learning path. Wang et al [46] propose a four phase learning portfolio mining approach, which use sequential

pattern mining, to extract learning features to create a decision tree which is used to predict which group a learner belongs to.

3.3.4 Casual data mining

Casual data mining, attempts to find whether one event (or observed construct) was the cause of another event (or observed construct), either by analyzing the covariance of two events or by using information about how one of the events was triggered. A casual relationship can be inferred, if a pedagogical event is randomly chosen using automated experimentation [27], and frequently leads to a positive learning outcome ragged.

3.4 Distillation of data for human judgment

The methods in this area of educational data mining are information visualization methods. However, the visualization used within EDM are often different than those often used for other information visualization problems [23], owing to the specific structure, and the meaning embedded within that structure, often present in educational data. These techniques facilitate to analyze large amount of information by representing the data in some visual display. Normally large qualities of data are represented or plotted as spreadsheet charts, scatter plot, 3D representations etc. The information visualized in statistical graph can be about assignment complement, admitted questions, exam score, etc. [39]. There are some specific visualization tools in educational data as CourseVis [24], visualizes data from a java on-line distance course inside WebCT. GISMO [25] uses Moodle students' tracking data as source data, and generates graphical representations that can be explored by course instructor. Listen tool [28] browses student-tutor interactions logs from project LISTEN's automated Reading Tutor, and generates graphical representations, which can be used by instructors to gain an understanding of their learners.

3.5 Discovery with models

In discovery with models, a model of a phenomenon is developed via prediction, clustering or knowledge engineering. This model is then used as a component in another analysis, such as relationship mining or prediction. Discovery with models has become an increasingly popular method in EDM research. Beck and Mostow [8], have supporting sophisticated analysis such as which learning material sub-categories of students will most benefit from. These methods can also be useful for finding, how variations in intelligent tutor design impacts students' behavior over time [22], and how different types of student behaviors impacts students' learning in different ways [13].

4. OTHER EDM METHODS

4.1 Outlier detection

Outlier detection discovers data points that are significantly different than the rest of data [19]. An outlier is an observation (or measurement) that is usually large or small when compared to other values in data set. In educational data mining, outlier analysis can be used to detect students with learning problems. Ueno [44] proposes a method of online outlier detection of learner's irregular learning processes by using the learners' response time data for e-learning data. Castro et al. [10] detects atypical behavior on the grouping structure of the users of virtual campus. Outlier detection also has been used for detecting regularities and deviations in the learners' or educators' actions with others [29].

4.2 Text mining

Text mining [17] methods can be viewed as an extension of data mining to text data and it is very related to web content mining. Text mining can work with semi-structured or unstructured datasets such as text documents, HTML files, emails etc. Ueno [43] uses data mining and text mining technologies for collaborative learning and a discussion board with evaluation between peers in an ILMS. Chen et al. [11] propose to automatically construct e-textbooks via web content mining. Tane et al. [41] use text mining and text clustering techniques in order to group documents according to their topics and similarities. The specific application of text mining techniques in e-learning can be used for: evaluating the progress of the thread discussion to see what the contribution to the topic is [15], identifying the main blocks of multimedia presentations and retrieve their internal properties [6], and personalized courseware construction proposed by Tang et al. [42].

4.3 Social Network Analysis (SNA)

Social network analysis is a field of study attempting to understand and measure relationships between entities in networked information. The social network analysis techniques and data mining techniques for information networks can be used to examine and assess online interactions [38]. The SNA uses the connection among units as data which relate them in system. In educational data mining, SNA can be used for mining group activities by analyzing the sociograms associated with a given group and the status of participants and the group cohesion of social interactions [34]. Rallo et al. [33] propose to use the data mining and social networks to interpret and analyze the structure and contents of online educational communities.

5. APPLICATIONS OF EDM METHODS

There have been a wide number of applications of educational data mining. The four areas of applications that have received the particular attentions, including individual learning from educational software, collaborative learning, computer-adaptive testing and analyzing students' performance.

One key area of applications is in improving student models. The model provides detailed information about a student's characteristics or states, such as knowledge motivation and attitude. The educational data mining methods have enabled researchers to make higher-level inferences about students' behavior, such as when a student is gaming the system [5], and when a student is engaging in self-explanation [40]. These students' models have increased the instructors' ability to predict the students' knowledge and future performance have enabled researchers to study what factors lead students to make specific choices in a learning setting.

A second key area of application is in discovering or improving models of the knowledge structure of the domain. In educational data mining, methods have been created for discovering accurate domain models directly from data. The combinations of psychometric modeling frameworks with advanced space-searching algorithms are posed as a prediction problem for the model discovery. Barnes [7] has developed algorithms which can automatically discover a Q-matrix from data. Pavlik et al. [31] has proposed algorithms for finding partial order knowledge structure models by looking at the covariance of individual item sets.

A third key area of application of EDM methods has been in studying the pedagogical support provided by learning software. Educational data miners are interested in discovering most effective pedagogical support. Learning decomposition fits exponential learning curves to performance data, relating students' success to the amount of each type of pedagogical support a student has received [8].

A fourth key area of application of educational data mining is for specific discovery about learning and learners. Discovery with models is a key method for scientific discovery via Educational data mining. Perera et al. [32] used the big five theory for teamwork as a driving theory to search for successful patterns of interaction within student teams.

6. CONCLUSION

The field of educational data mining is a ripe for explosive growth. Machine learning and data mining have developed a vast array of tools and techniques that have been well-studied and examined in myriad context. Educational data mining can be applied to wide areas of research including e-learning, intelligent tutoring systems, text mining, social network mining etc. Educational data mining is a very recent area, there is an important number of contributions published in journals, international congress, specific workshops like international Journal of Artificial Intelligence in education (IJAIED), International Conference of Intelligent Tutoring System (ITS), International Conference on Educational data mining (EDM), Journal of Educational data mining (JEDM), International conference on data Mining for user Modeling (DMUM) etc.

In education, EDM can function as a replacement for less accurate but more established psychometric techniques. Educational data mining is an interactive cycle of hypothesis formation, testing and refinements that alternates between two complementary types of activities. One type of activity is qualitative analysis, focuses on understanding individual tutorial events. Other type involve, knowledge tracing analyses the growth curve by aggregating over successive opportunities to apply skills.

7. REFERENCES

[1] Agarwal R., & Srikant R., "Mining sequential patterns", In Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan, pp. 3–14, 2005.

[2] Agarwal R., Imielinski T., & Swami A., "Mining association rules between sets of items in large databases", In Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, USA, pp. 1–22, 1993.

[3] Amershi S., Conati C., "Automatic Recognition of Learner Groups in Exploratory Learning Environments", Proceedings of ITS 2006, 8th International Conference on Intelligent Tutoring Systems, 2006.Tavel, P. Modeling and Simulation Design. AK Peters Ltd, 2007.

[4] Baker R.S.J.D., "Data Mining For Education. In International Encyclopedia of Education (3rd edition)", B. MCGAW, PETERSON, P., BAKER Ed. Elsevier, Oxford, UK, 2009.Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, pp. 1289-1305, 2003.

[5] Baker R.S.J.D., "Is Gaming the System State-or-Trait? Educational Data Mining Through the Multi-Contextual Application of a Validated Behavioral Model", Complete On-Line Proceedings of the Workshop on Data Mining

for User Modeling at the 11th International Conference on User Modeling, pp. 76-80, 2007.

[6] Bari M., & Benzater B., "Retrieving data from pdf interactive multimedia productions", In International conference on human system learning: Who is in control? pp. 321–330, 2005.

[7] Barnes T., "The q-matrix method: Mining student response data for knowledge", In Proceedings of the AAAI-2005 Workshop on Educational Data Mining, 2005.

[8] Beck J.E. and Mostow J., "How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students", In Proceedings of the 9th International Conference on Intelligent Tutoring Systems, pp. 353-362, 2008.

[9] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students' Performance", In International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, pp. 63-69, 2011.

[10] Castro F., Vellido A., Nebot A., & Minguillon J., "Detecting atypical student behavior on an e-learning system", In I Simposio Nacional de Tecnologas de la Informacin y las Comunicaciones en la Educacin, Granada, pp. 153–160, 2005.

[11] Chen J., Li Q., Wang L., & Jia W., "Automatically generating a textbook on the web", In International conference on advances in web-based learning, Beijing, China, pp. 35–42, 2004.

[12] Choquet C., Luengo V. and Yacef K., Proceedings of "Usage Analysis in Learning Systems" workshop, held in conjunction with AIED 2005, Amsterdam, The Netherlands, 2005.

[13] Cocea M., Hershkovitz A. and Baker R.S.J.D., "The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate?", In Proceedings of the 14th International Conference on Artificial Intelligence in Education, pp. 507-514, 2009.

[14] Dempster A., Larid N., Rubin D., "Maximum Likelihood estimation from incomplete data via EM Algorithm", Journal of Royal Statistics Society, 39(1), pp. 1-38, 1997.

[15] Dringus L., & Ellis T., "Using data mining as a strategy for assessing asynchronous discussion forums", Computer & Education Journal, 45, pp. 141–160, 2005.

[16] Duda R. O., Hart P. E., & Stork D. G., "Pattern classification", Wiley Interscience, 2000.

[17] Grobelnik M., Mladenic D., & Jermol M., "Exploiting text mining in publishing and education", In Proceedings of the ICML workshop on data mining lessons learned, Sydney, Australia, pp. 34–39, 2002.

[18] Han J. W., Kamber M., "Data Mining: Concepts and Techniques", 2nd Edition. The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2006.

[19] Hodge V., & Austin J., "A survey of outlier detection methodologies", Artificial Intelligence Review, 22(2), pp. 85–126, 2004.

[20] Hwang G. J., Hsiao C. L., & Tseng C. R., "A computer-assisted approach to diagnosing student learning problems in science courses", Journal of Information Science and Engineering, 19, pp. 229–248, 2003.

[21] Jain A. K., Murty M. N., & Flynn P. J., "Data clustering: A review", ACM Computing Surveys, 31(3), pp. 264–323, 1999.

[22] Jeong H. and Biswas G., "Mining Student Behavior Models in Learning-by-Teaching Environments", In Proceedings of the 1st International Conference on Educational Data Mining, pp. 127-136, 2008.

[23] Kay J., Maisonneuve N., Yacef K. and Reimann P., "The Big Five and Visualizations of Team Work Activity", In Intelligent Tutoring Systems, M. IKEDA, K.D. ASHLEY and T.-W. CHAN Eds. Springer-Verlag, Taiwan, pp. 197-206, 2006.

[24] Mazza R., & Dimitrova V., "Visualizing student tracking data to support instructors in web-based distance education", In International World Wide Web conference, New York, USA, pp. 154–161, 2004.

[25] Mazza R., & Milani C., "Exploring usage analysis in learning systems: Gaining insights from visualizations", In Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education. New York, USA, pp. 1–6, 2005.

[26] Merceron A., & Yacef K., "Mining student data captured from a web-based tutoring tool: Initial exploration and results", Journal of Interactive Learning Research, 15(4), pp. 319–346, 2004.

[27] Mostow J., "Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods", In C. K. Kinzer & L. Verhoeven (Eds.), Interactive Literacy Education: Facilitating Literacy Environments through Technology, pp. 117-148, 2008.

[28] Mostow J., Beck J., Cen H., Cuneo A., Gouvea E., & Heiner C., "An educational data mining tool to browse tutor–student interactions: Time will tell!", In Proceedings of the workshop on educational data mining, Pittsburgh, USA, pp. 15–22, 2005.

[29] Muehlenbrock M., "Automatic action analysis in an interactive learning environment", In Proceedings of the workshop on usage analysis in learning systems at the 12th international conference on artificial intelligence in education, Amsterdam, The Netherlands, pp.73–80, 2005.

[30] Pahl C., & Donnellan C., "Data mining technology for the evaluation of web-based teaching and learning systems", In Proceedings of the Congress E-learning. Montreal, Canada, pp. 1–7, 2003.

[31] Pavlik P., Cen H., Wu L. and Koedinger K., "Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor", In Proceedings of the 1st International Conference on Educational Data Mining, pp. 77-86, 2008.

[32] Perera D., Kay J., Koprinska I., Yacef K. and Zaiane O., "Clustering and sequential pattern mining to support team learning", IEEE Transactions on Knowledge and Data Engineering 21, pp. 759-772, 2009.

[33] Rallo R., Gisbert M., & Salinas J., "Using data mining and social networks to analyze the structure and content of educative online communities", In International conference on multimedia and ICTs in education, Caceres, Spain, pp. 1–10, 2005.

[34] Reyes P., & Tchounikine P., "Mining learning groups' activities in forum-type tools", In Proceedings of the 2005 conference on computer support for collaborative learning, Taiwan, pp 509–513, 2005.

[35] Romero C., & Ventura S., "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, 33(1), pp. 135–146, 2007.

[36] Romero C., Ventura S., Bra P., & Castro C., "Discovering prediction rules in aha! Courses", In Proceedings of the International Conference on User Modeling, pp. 25–34, 2003.

[37] Romero C., Ventura S., Espejo P.G. and Hervas C., "Data Mining Algorithms to Classify Students", In Proceedings of the 1st International Conference on Educational Data Mining, pp. 8-17, 2008.

[38] Scott J., "Social network analysis: A handbook", (2nd ed.). Newberry Park, CA: Sage, 2000.

[39] Shen R., Yang F., & Han P., "Data analysis center based on learning platform", In Proceedings of the 5th international workshop on the internet challenge: Technology and applications, pp. 19–28, 2002.

[40] Shih B., Koedinger K.R., Scheines R., "A Response-Time Model for Bottom-Out Hints as Worked Examples", Proceedings of the First International Conference on Educational Data Mining, pp. 117-126, 2008.

[41] Tane J., Schmitz C., & Stumme G., "Semantic resource management for the web: An e-learning application", In Proceedings of the WWW conference, New York, USA, pp. 1–10, 2004.

[42] Tang C., Yin H., Li T., Lau R., Li Q., & Kilis D., "Personalized courseware construction based on web data mining", In Proceedings of the first international conference on web information systems engineering, Washington, DC, USA, pp. 204–211, 2000.

[43] Ueno M., "Data mining and text mining technologies for collaborative learning in an ILMS samurai'', In IEEE international conference on advanced learning technologies, Joensuu, Finland, pp. 1052–1053, 2004a.

[44] Ueno M., "Online outlier detection system for learning time data in e-learning and its evaluation", In International conference on computers and advanced technology in education, Beijing, China, pp. 248–253, 2004b.

[45] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", American Association for Artificial Intelligence, pp. 37-54, 1997.

[46] Wang W., Weng J., Su J., & Tseng S., "Learning portfolio analysis and mining in SCORM compliant environment", In ASEE/IEEE frontiers in education conference, pp. 17–24. 2004.

[47] Witten I.H., Frank E., "Data Mining: Practical Machine Learning Tools and Techniques", 2nd Edition. San Francisco: Morgan Kaufmann, 2005.

[48] www.educationaldatamining.org.

[49] Yu P., Own C., & Lin L., "On learning behavior analysis of web based interactive environment", In Proceedings of the implementing curricular change in engineering education, Oslo, Norway, pp. 1–10, 2001.

[50] Zaïane O., & Luo J., "Web usage mining for a better web-based learning environment", In Proceedings of conference on advanced technology for education, Banff, Alberta, pp.60–64, 2001.