Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 135–146
Expert Systems with Applications
www.elsevier.com/locate/eswa

Educational data mining: A survey from 1995 to 2005

C. Romero *, S. Ventura

Department of Computer Sciences, University of Cordoba, Cordoba, Spain

Abstract

Currently there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community. This paper surveys the application of data mining to traditional educational systems, particular web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems. Each of these systems has different data source and objectives for knowledge discovering. After preprocessing the available data in each case, data mining techniques can be applied: statistics and visualization; clustering, classification and outlier detection; association rule mining and pattern mining; and text mining. The success of the plentiful work needs much more specialized work in order for educational data mining to become a mature area.

Keywords: Data mining; Educational systems; Web mining; Web-based educational systems

1. Introduction

During the past decades, the most important innovations in educational systems are related to the introduction of new technologies (Ha, Bae, & Park, 2000) as web-based education. This is a form of computer-aided instruction virtually independent of a specific location and any specific hardware platform (Brusilovsky & Peylo, 2003). It has considerably gained in importance and thousands of web courses have been deployed in the past few years. But many of the current web-based courses are based on static learning materials, which do not take into account the diversity of students. Adaptive and intelligent web-based educational systems have been seen as a solution to individually richer learning environments. These systems try to offer learners personalized education by building a model of the individual's goals, preferences, and knowledge. Data mining or knowledge discovery in databases (KDD) is the automatic extraction of implicit and interesting patterns from large data collections (Klosgen & Zytkow, 2002). KDD can be used not only to learn the model for the learning process (Hamalainen, Suhonen, Sutinen, & Toivonen, 2004) or student modeling (Tang & McCalla, 2002) but also to evaluate and to improve e-learning systems (Zaïane & Luo, 2001) by discovering useful learning information from learning portfolios (Hwang, Chang, & Chen, 2004).

In conventional teaching environments, educators are able to obtain feedback on student learning experiences in face-to-face interactions with students, enabling a continual evaluation of their teaching programs (Sheard, Ceddia, Hurst, & Tuovinen, 2003). Decision making of classroom processes involves observing a student's behavior, analyzing historical data, and estimating the effectiveness of pedagogical strategies. However, when students work in electronic environments, this informal monitoring is not possible; educators must look for other ways to attain this information. Organizations, which run distance education sites, collect large volumes of data, automatically generated by web servers and collected in server access logs. Web-based learning environments are able to record most learning behaviors of the students, and are hence able to provide a huge amount of learning profile. Recently, there is a growing interest in the automatic analysis of learner interaction data with web-based learning environments

* Corresponding author. Tel.: +34 957 212172; fax: +34 957 218630.
E-mail address: cromero@uco.es (C. Romero).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.04.005

(Muehlenbrock, 2005). In order to provide a more effective learning environment, data mining techniques can be applied (Ingram, 1999). Data mining is a step in the overall process of KDD that consists of preprocessing, data mining and postprocessing. Data mining has already been successfully applied in e-commerce (Srivastava, Cooley, Deshpande, & Tan, 2000), and it has begun to be used in e-learning with promising results. Although the discovery methods used in both areas (e-commerce and e-learning) are similar (Hanna, 2004), there are some important differences between them:

- *Domain*. The e-commerce purpose is to guide clients in purchasing while the e-learning purpose is to guide students in learning (Romero, Ventura, & Bra, 2004).
- *Data*. In e-commerce the used data are normally simple web server access log, but in e-learning there is more information about a student's interaction (Pahl & Donnellan, 2003). The user model is also different in both systems.
- *Objective*. The objective of data mining in e-commerce is increasing profit, that is tangible and can be measured in terms of amounts of money, number of customers and customer loyalty. And the objective of data mining in e-learning is to improving the learning. This goal is more subjective and more subtle to measure.
- *Techniques*. Educational systems have special characteristics that require a different treatment of the mining problem. As a consequence, some specific data mining techniques are needed to address in particular the process of learning (Li & Zaïane, 2004; Pahl & Donnellan, 2003). Some traditional techniques can be adapted, some cannot.

The application of knowledge extraction techniques to educational systems in order to improve learning can be viewed as a formative evaluation technique. Formative evaluation (Arruabarrena, Pérez, López-Cuadrado, & Vadillo, 2002) is the evaluation of an educational program while it is still in development, and with the purpose of continually improving the program. Examining how students use the system is one way to evaluate the instructional design in a formative manner and it may help the educator to improve the instructional materials (Ingram, 1999). Data mining techniques can discover useful information that can be used in formative evaluation to assist educators establish a pedagogical basis for decisions when designing or modifying an environment or teaching approach. The application of data mining in educational systems is an iterative cycle of hypothesis formation, testing, and refinement (see Fig. 1). Mined knowledge should enter the loop of the system and guide, facilitate and enhance learning as a whole. Not only turning data into knowledge, but also filtering mined knowledge for decision making.

As we can see in Fig. 1, educators and academics responsible are in charge of designing, planning, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage and interaction, different data mining techniques can be applied in order to discover useful knowledge that helps to improve the e-learning process. The discovered knowledge can be used not only by providers (educators) but also by own users (students). So, the application of data mining in educational systems can be oriented to different actors with each particular point of view (Zorrilla, Menasalvas, Marin, Mora, & Segovia, 2005):

- *Oriented towards students* (Heraud, France, & Mille, 2004; Farzan, 2004; Lu, 2004; Tang & McCalla, 2005; Zaïane, 2002). The objective is to recommend to learners activities, resources and learning tasks that would favour and improve their learning, suggest good learning experiences for the students, suggest path pruning and shortening or simply links to follow, based on the tasks already done by the learner and their successes, and on tasks made by other similar learners, etc.

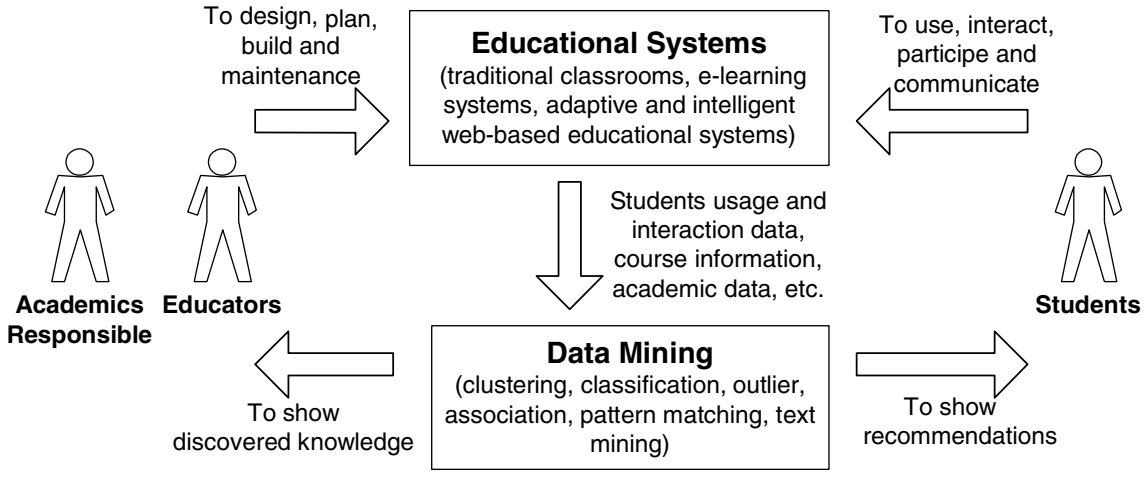

Fig. 1. The cycle of applying data mining in educational systems.

- *Oriented towards educators* (Ha et al., 2000; Hamalainen et al., 2004; Merceron & Yacef, 2004; Minaei-Bidgoli & Punch, 2003; Mor & Minguillon, 2004; Muehlenbrock, 2005; Pahl & Donnellan, 2003; Romero et al., 2004; Silva & Vieira, 2002; Talavera & Gaudioso, 2004; Tang et al., 2000; Ueno, 2004b; Zaïane & Luo, 2001). The objective is to get more objective feedback for instruction, evaluate the structure of the course content and its effectiveness on the learning process, classify learners into groups based on their needs in guidance and monitoring, find learning learner's regular as well as irregular patterns, find the most frequently made mistakes, find activities that are more effective, discover information to improve the adaptation and customization of the courses, restructure sites to better personalize courseware, organize the contents efficiently to the progress of the learner and adaptively constructing instructional plans, etc.
- *Oriented towards academics responsible and administrators* (Becker, Ghedini, & Terra, 2000; Grob, Bensberg, & Kaderali, 2004; Luan, 2002; Ma, Liu, Wong, Yu, & Lee, 2000; Peled & Rashty, 1999; Sanjeev & Zytkow, 1995; Urbancic, Skrjanc, & Flach, 2002). The objective is to have parameters about how to improve site efficiency and adapt it to the behavior of their users (optimal server size, network traffic distribution, etc.), have measures about how to better organize institutional resources (human and material) and their educational offer, enhance educational programs offer and determine effectiveness of the new computer mediated distance learning approach.

There are many general data mining tools that provide mining algorithms, filtering and visualization techniques. Some examples of commercial and academic tool are DBMiner, Clementine, Intelligent Miner, Weka, etc. (Klosgen & Zytkow, 2002). However these tools are not specifically designed and maintained for pedagogical purposes and it is cumbersome for an educator who does not have an extensive knowledge in data mining to use these tools (Zaïane, Xin, & Han, 1998). In order to solve this problem, some specific educational data mining, statistical and visualization tools have been developed to help educators in analyzing the different aspects of the learning process (see Table 1).

Table 1
Some specific educational data mining, statistics and visualization tools

Tool name	Authors	Mining task
Mining tool	Zaïane and Luo (2001)	Association and patterns
MultiStar	Silva and Vieira (2002)	Association and classification
Data Analysis Center	Shen et al. (2002)	Association and classification
EPRules	Romero et al. (2003)	Association
KAON	Tane et al. (2004)	Text mining and clustering
TADA-ED	Merceron and Yacef (2005)	Classification and association
O3R	Becker et al. (2005)	Sequential patterns
Synergo/ColAT	Avouris et al. (2005)	Statistics and visualization
GISMO/CourseVis	Mazza and Milani (2005)	Visualization
Listen tool	Mostow et al. (2005)	Visualization
TAFPA	Damez et al. (2005)	Classification
iPDF_Analyzer	Bari and Benzater (2005)	Text mining

We have divided this paper into the following sections. We first review some different types of educational systems and how data mining can be applied in each of them. We then describe the data mining techniques that have been applied in educational systems grouping them by task. Finally, we summarize the main conclusions and we draw some future research.

2. Educational systems: data and objectives

Data mining can be applied to data coming from two types of educational systems: traditional classroom and distance education. It is necessary to deal separately with the application of data mining techniques in each type due to the fact that they have different data sources and objectives.

2.1. Traditional classrooms

Traditional classroom environments are the most widely used educational systems. It is based on face-to-face contact between educators and students organized through lecturers. There are a lot of different subtypes: private and public education, elementary and primary education, adult education, higher, tertiary and academic education, special education, etc. They have been criticized because they encourage passive learning, ignore individual differences and needs of the learners, and do not pay attention to problem solving, critical thinking, or other higher order thinking skills (Johnson, Arago, Shaik, & Palma-Rivas, 2000). In conventional classrooms, educators attempt to enhance instructions by monitoring student's learning processes and analyzing their performances by paper records and observation. They can also use information about student attendance, course information, curriculum goals, and individualized plan data. And educational institution has many diverse and varied sources of information (Ma et al., 2000): traditional databases (with a student's information, educator's information, class and schedule information, etc.), online information (online web pages and course content pages), multimedia databases, etc.

Data mining can help each actor of the learning process. Institutions would like to know which students will enroll in a particular course and which students will need assistance in order to graduate. An administrator may wish to find out information such as the admission requirements and to predict the class enrollment size for timetabling. Students may wish to know how best to select courses

based on prediction of how well they will perform in the courses selected. Instructors may wish to know what learning experiences are most contributive to overall learning outcomes, why is one class outperforming the other, similar groups of students, etc. There are some works about the application of data mining in traditional education. One of the first articles about the use of data mining in education to understand the student enrollment was written by Sanjeev and Zytkow (1995). They apply knowledge discovery in the form of statements "Pattern P holds for data in Range R" to university databases. The results were presented to a senior university administrator in order to make strategic decisions about the institutional policies. Another work on the use of KDD to identity and understand whether curriculum revisions can affect students in a Brazilian university was done by Becker et al. (2000). They verify the qualitative impact of revisions and evaluate it using a number of techniques, such as summarization, association, classification. In a related work, the objective is to select the weak students to attend remedial classes (Ma et al., 2000). They use a scoring function that is based on association rules. First, they identify the potential weak students and then select the course that each weak student is recommended to take. Finally, an application in higher education for doing a comprehensive analysis of student characteristics is done by Luan (2002). He proposes to use different unsupervised (Kohonen nets) and supervised (C5.0, genetic algorithms, etc.) data mining algorithms to do clustering and prediction in order to enable educational institutions to better allocate resources and staff, proactively manage student outcomes, and improve the effectiveness of alumni development.

2.2. Distance education

Distance education or distance learning consists of techniques and methods providing access to educational programs for students who are separated by time and space from lecturers. e-Learning systems lack a closer student–educator relationship (one to one). There are different subtypes of distance education: paper-based correspondence education, videotape education, computer-aided education (multimedia education, internet education or web-based education), etc. Currently, the most used is web-based education allowing students to conveniently learn via the Internet. Web-based education is a form of distance education delivered over the Internet (Johnson et al., 2000). Today, there are a lot of terms used to refer to web-based education such as e-learning, e-training, online instruction, web-based learning, web-based training, web-based instruction, etc. And there are different types of web-based systems: synchronous and asynchronous, collaborative and non-collaborative, closed corpus and open corpus, etc. These web-based education systems can normally record the student's accesses in web logs that provide a raw trace of the learners' navigation on the site. There are several types of logs (Srivastava et al., 2000):

- *Server log file*. This constitutes the most widely used data source for performing data mining, containing just the bare details of timing, path, and input-response. There are a variety of formats, such as common log format (CLF), extended log format (ELF), etc. (Koutri, Avouris, & Daskalaki, 2004). Normally, there is a single log file for all students.
- *Client log file*. This consists of a set of log files, one per student, and contains information about the interaction of the user with the system. Can be implemented by a remote agent (such Javascripts, Java Applets), modifying the source code of an existing browser, or using cookies.
- *Proxy log file*. This consists of a set of log files of caching between client browsers and web servers. This information complements server log file information.

It should be noted that the concept of logging may include restrictions by law. Therefore, whenever a log system authenticates users it should not relate to a person's true identity but primarily they as individual persons (Rahkila & Karjalainen, 1999). Log files also have several inherent limitations, tracking for files not users, simple clicks and not learning activities, not capturing contextual information, recognizing specific computers not specific people, having incomplete and incorrect information problem, and some technical aspects of web browser (as the cache) may prevent to record logs. To address these problems, authors have proposed several solutions. Yu, Own, and Lin (2001) propose to use another way to record a learner's portfolio that includes the learning path, preferred learning course, grade of course, and learning time, etc. Li and Zaïane (2004) use more information channels to model user navigational behavior: web access logs, the structure of a visited web site, and the content of visited web pages. Avouris, Komis, Fiotakis, Margaritis, and Voyiatzaki (2005) expand automatically generated log files by introducing contextual information as additional events and by associating comments and static files. Monk (2005) combines data on the activity with content and user profiles in a composite information model. Ingram (1999) combines data with other inquiry methods, such as informal chatting with students, in class shows of hands, surveys, or written feedback about the web site. Iksal and Choquet (2005) propose to use a specific usage tracking meta-language to describe the track semantics recorded by web-based educational systems. Markham et al. (2003) propose to use software agents to extract data from the e-learning environment and to organize them in intelligent ways.

Next, we distinguish between three different types of web-based education systems: particular web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems.

2.2.1. Particular web-based courses

Particular web-based courses are specific courseware that use standard HTML (HyperText Markup Language).

There are a lot of courses, tutorials, etc. of this type on the Internet, and as another web site, they have the same kinds of data sources (Srivastava et al., 2000):

- *Content:* The real data in the web pages, i.e. the data the web page were designed to convey to the users. This usually consists of texts, graphics, videos, sounds, etc.
- *Structure:* Data which describe the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the HTML tag becomes the root of the tree. The principal kind of inter-page structure information is hyper-links connecting one page to another.
- *Usage:* Data that describe the pattern of usage of web pages. There are two main types of students' information (Silva & Vieira, 2002): information about the student's actions and communications, and information about the student's activities in the course.
- *User profile:* Data that provide demographic information about users of the web site. This includes registration data and customer profile information.

Data mining can be used to know how students use the course, how a pedagogical strategy impacts different types of students, in which order the students study subtopics, what are the pages/topics that students skip, how much time the students spend with a single page, a chapter or the full course, etc.

2.2.2. Well-known learning content management systems

Well-known learning content management systems (LCMS) are platforms that offer a great variety of channels and workspaces to facilitate information sharing and communication between participants in a course, let educators distribute information to students, produce content material, prepare assignments and tests, engage in discussions, manage distance classes and enable collaborative learning with forums, chats, file storage areas, news services, etc. Some examples of commercial LCMS are Blackboard, Virtual-U, WebCT, TopClass, etc. and some example of free LCMS are Moodle, Ilias, Claroline, aTutor, etc. (Paulsen, 2003). These systems accumulate large log data of the students' activities and usually have built-in student monitoring features (Mazza & Milani, 2005). They can record whatever student activities it involves, such as reading, writing, taking tests, performing various tasks in real or virtual environments, even communicating with peers (Mostow, 2004). They normally also provide a database that stores all the systems information: personal information of the users (profile), academic results, user's interaction data, etc.

Although some platforms offer reporting tools, when there is a great number of students, it becomes hard for a tutor to extract useful information. Data mining can be applied to explore, visualize and analyze data in order to identify useful patterns (Talavera & Gaudioso, 2004) and to evaluate web activity to get more objective feedback for your instruction and knowing more about how the students learning on the LCMS (Zaïane & Luo, 2001).

2.2.3. Adaptive and intelligent web-based educational systems

Adaptive and intelligent web-based educational systems (AIWBES) provide an alternative to the traditional just-put-it-on-the-web approach in the development of web-based educational courseware (Brusilovsky & Peylo, 2003). AIWBES attempt to be more adaptive by building a model of the goals, preferences and knowledge of each individual student and using this model throughout the interaction with the student in order to adapt to the needs of that student. AIWES are the result of a joint evolution of intelligent tutoring systems (ITS) and adaptive hypermedia systems (AHS). Some examples of ITS are SQL-Tutor, German Tutor, ActiveMath, VC-Prolog-Tutor, and some examples of AHS are AHA!, InterBook, KBS-Hyperbook, WebCOBALT (Brusilovsky & Peylo, 2003). The data from AIWEBS are semantically richer and can lead to more diagnostic analysis than data from traditional web-based education system (Merceron & Yacef, 2004). The available data come from the domain model (which may be structured into an ontology), pedagogical dataset (set of problems, their answer and complexity information), interaction log files (data related with user interaction) and student model (list of the satisfactions and violations constraints). AIWBES use a standard student model (used internally by the tutoring system), but, for the purpose of data mining, it is necessary to have a new model of student interaction with augmented information with contextual data. These student's interaction can be analyzed at a number of different layers of granularity: course, sessions, problems, attempts and constraints (Nilakant & Mitrovic, 2005).

Data mining can be used in order to know the causes of problems in the system, for example, incorrect feedback statements (Nilakant & Mitrovic, 2005), to adapt the level to the progress of the learner (Romero et al., 2004), to suggest personalized learning experiences and activities for the students (Tang & McCalla, 2005).

3. Data preprocessing

Data preprocessing allows to transform the original data into a suitable shape to be used by a particular mining algorithm. So, before applying the data mining algorithm, a number of general data preprocessing tasks have to be addressed (Koutri et al., 2004; Zorrilla et al., 2005):

- *Data cleaning*. It is one of the major preprocessing tasks, to remove irrelevant items and log entries that are not needed for the mining process such us graphics, scripts.
- *User identification*. Process of associating page references to the connected user.
- *Session identification*. It takes all of the page references for a given user and course in a log and breaks them up into user sessions. In our particular case, we have

initially considered a new session when a change in a user course happens or when the time interval between two successive inter-transaction clicks ups 30 min (Zorrilla et al., 2005).
- *Path completion*. It fills in page references that are missing due to browser and proxy server caching.
- *Transaction identification*. It breaks down sessions into smaller units, referred to as transactions or episodes.
- *Data transformation and enrichment*. It consists of calculating new attributes from the existing ones, conversing of numerical attributes into nominal attributes, providing meaning to references contained in the log, etc.
- *Data integration*. It is the integration and synchronization of data from heterogeneous sources.
- *Data reduction*. It is for reducing data dimensionality.

Additionally, data preprocessing of web-based educational systems has some specific issues:

- Most of the systems use user authentication (password protection) in which logs have entries identified by users since the users have to log-in, and sessions are already identified since users may also have to log-out (Rahkila & Karjalainen, 1999).
- Most of the systems record the students' interactions not only in log files but also directly in databases. If this is not the case, during the preparation process, data for each individual student (profile, logs, etc.) can be aggregated to a database (Talavera & Gaudioso, 2004). Databases are more powerful than typical log text files and provide an analysis easier, more flexible and less bug prone.
- Data transformation is more oriented to a better interpretation of data. Numerical values are discretized or transformed into ranges that provide a much more comprehensible view of the data. New attributes result from other current attributes in a specific attribute derivation. The derivation performs some kind of aggregation, for example, each attempt is grouped into a new number of attempt attribute (Nilakant & Mitrovic, 2005).
- In the division of individual visit sessions into transactions, can be identified subsessions or missions with coherent information needs in which the identified sequence is based on the real content of pages (Li & Zaïane, 2004). Besides different meanings of interaction at different levels of abstraction can be distinguished (Pahl & Donnellan, 2003): learning and training interaction, human–computer interface and multimedia and service interaction.
- The data filtration uses specific educational concepts as number of attempts, number of repeated reading, level of knowledge, etc. Normally, data is filtered by defining some condition on one or more attributes and removing the instances that violate it (Nilakant & Mitrovic, 2005).

The educators have to actively participate in the previous preprocessing task, for example, indicating specific data filtration and attribute derivation or transformation, etc. So, it is needed to enhance preprocessing facilities that prepare the e-learning data in a meaningful and useful way.

4. Data mining techniques in educational systems

Data mining is a multidisciplinary area in which several computing paradigms converge: decision tree construction, rule induction, artificial neural networks, instance-based learning, Bayesian learning, logic programming, statistical algorithms, etc. (Klosgen & Zytkow, 2002). Next, we are going to describe some specific application of data mining techniques grouped by tasks, in web-based educational systems (see Table 2).

4.1. Statistics and visualization

Student's usage statistics are often the starting point of evaluations of an e-learning system, although they are usually not considered as data mining techniques (Tsantis & Castellani, 2001). Formal statistical inference is assumption driven in the sense that a hypothesis is formed and then tested against the data. Data mining, in contrast, is discovery driven in the sense that the hypothesis is automatically extracted from the data.

Usage statistics may be extracted using standard tools designed to analyze web server logs as AccessWatch, Analog, Gwstat, WebStat, etc. (Zaïane et al., 1998). But there are some specific statistical tools in educational data as Synergo/ColAT (Avouris et al., 2005). Some example of usage statistics are simple measures such as the total number of visits and number of visits per page (Pahl & Donnellan, 2003). Some other general statistics show the connected learner distribution over time, the most frequent acceded courses, how learners establish many learning sessions over time (Zorrilla et al., 2005). Besides, some specific statistical in AIWBES can show the average number of constraint violations, the average problem complexity, the total time spent in attempts (Nilakant & Mitrovic, 2005). More complex statistical tests of procedures such as regression analysis, correlation analysis, multivariate statistical methods. (Zarzo, 2003) need to use a more powerful statistical tools as SPSS, SAS, S, R, Statistica, etc. (Klosgen & Zytkow, 2002). If data are stored in a relational database, then SQL queries (Heiner, Beck, & Mostow, 2004; Merceron & Yacef, 2003) can provide functionality for a number of simple statistical operations such as standard deviation, mode, sample size, etc. (Nilakant & Mitrovic, 2005). But the information obtained from usage statistics is not always easy to interpret to the educators and then other techniques have to be used.

Information visualization techniques can be used to graphically render complex, multidimensional student tracking data collected by web-based educational systems (Mazza & Milani, 2005). These techniques facilitate to analyze large amounts of information by representing the data in some visual display. Normally large quantities of

Table 2
Works about applying data mining techniques in educational systems

Authors	Mining task	Educational system
Sanjeev and Zytkow (1995)	Sequence pattern	Traditional education
Zaïane et al. (1998)	Statistic and sequence pattern	LCM systems
Beck and Woolf (2000)	Prediction	AIWBE system
Becker et al. (2000)	Association and classification	Traditional education
Chen et al. (2000)	Classification	Web-based course
Ha et al. (2000)	Association	Web-based course
Ma et al. (2000)	Association	Traditional education
Tang et al. (2000)	Text mining	AIWBE system
Yu et al. (2001)	Association	Web-based course
Zaïane and Luo (2001)	Sequence pattern	LCM system
Luan (2002)	Clustering and prediction	Traditional education
Pahl and Donnellan (2003)	Sequence pattern and statistics	LCM system
Shen et al. (2002)	Visualization	LCM system
Wang (2002)	Association and sequence pattern	Web-based course
Merceron and Yacef (2003)	Statistic	AIWBE system
Minaei-Bidgoli and Punch (2003)	Classification	Web-based course
Shen et al. (2003)	Sequence pattern and clustering	Web-based course
Zarzo (2003)	Statistic	Web-based course
Arroyo et al. (2004)	Prediction	AIWBE system
Baker et al. (2004)	Classification	AIWBE system
Chen et al. (2004)	Text mining	Web-based course
Freyberger et al. (2004)	Association	AIWBE system
Hamalainen et al. (2004)	Classification	AIWBE system
Heiner et al. (2004)	Statistic	AIWBE system
Lu (2004)	Association	AIWBE system
Merceron and Yacef (2004)	Association	AIWBE system
Minaei-Bidgoli et al. (2004)	Association	Web-based course
Mor and Minguillon (2004)	Clustering	LCM system
Romero et al. (2004)	Association	AIWBE system
Talavera and Gaudioso (2004)	Clustering	LCM system
Ueno (2004b)	Outlier detection	Web-based course
Ueno (2004a)	Text mining	Web-based course
Wang et al. (2004)	Sequence pattern and clustering	LCM system
Li and Zaïane (2004)	Association	LCM system
Avouris et al. (2005)	Statistic	Web-based course
Castro et al. (2005)	Outlier detection	LCM system
Dringus and Ellis (2005)	Text mining	LCM system
Feng et al. (2005)	Prediction	AIWBE system
Hammouda and Kamel (2005)	Text mining	Web-based course
Markellou et al. (2005)	Association	Web-based course
Mazza and Milani (2005)	Visualization	LCM system
Mostow et al. (2005)	Visualization	AIWBE system
Muehlenbrock (2005)	Outlier detection	AIWBE system
Nilakant and Mitrovic (2005)	Statistic	AIWBE system
Tang and McCalla (2005)	Clustering	AIWBE system
Zorrilla et al. (2005)	Statistic	LCM system
Damez et al. (2005)	Classification	AIWBE system
Bari and Benzater (2005)	Text mining	LCM system

raw instance data are represented or plotted as spreadsheet charts, scatterplot, 3D representations, etc. The information visualized in statistical graphs can be about assignment complement, admitted question, exam score, etc. (Shen, Yang, & Han, 2002). Visualization techniques have been used to visualize social aspects in computer-supported collaborative learning, community relationships in peer-to-peer systems, and conversations in online groups. Instructors can manipulate the graphical representations generated, which allow them to gain an understanding of their learners and become aware of what is happening in distance classes. There are some specific visualization tools in educational data as GISMO/CourseVis (Mazza & Milani, 2005) and Listen tool (Mostow et al., 2005).

4.2. Web mining

Web mining (Srivastava et al., 2000) is the application of data mining techniques to extract knowledge from web data. There are three main web mining categories from the used data viewpoint: web content mining is the process of extracting useful information from the contents of web documents; web structure mining is the process of discovering structure information from the web; and web usage mining (WUM) that is the discovering of meaningful patterns from data generated by client–server transactions on one or more web localities. But there are two types of web mining categories from the used system viewpoint (Li & Zaïane, 2004): offline web mining, that is used to discover patterns and other useful information to help educators to validate the learning models and restructure the web site; and online or integrated web mining in which the patterns automatically discovered are fed into an intelligent software system or agent that could assist learners in their online learning endeavours. The mined patterns are used on-the-fly by the system to improve the application or its functions.

There are different web mining techniques applied to educational systems, but almost all of them can be grouped in one of the three next ones: clustering, classification and outlier detection; association rule mining and sequential pattern mining; and text mining.

4.2.1. Clustering, classification and outlier detection

Clustering is a process of grouping physical or abstract objects into classes of similar objects. Clustering and classification (Klosgen & Zytkow, 2002) are both classification methods. Clustering is an unsupervised classification and classification is a supervised classification. Classification and prediction are also related techniques. Classification predicts class labels, whereas prediction predicts continuous-valued functions. On the other hand, an outlier is an observation (or measurement) that is unusually large or small relative to the other values in a dataset. Outliers typically are attributable to one of the following causes: the measurement is observed, recorded, or entered into the computer incorrectly; the measurements come from a different population; the measurement is correct, but represents a rare event.

All these methods have been applied to web-based educational systems. Clustering can group together a set of pages with similar contents, users with similar navigation behavior or navigation sessions. Classification allows characterizing the properties of a group of user profiles, similar pages or learning sessions. And outlier detection can detect students with learning problems. Next, we describe some works about the application of these techniques in different types of web-based educational systems:

- *Particular web-based courses*. Chen, Liu, Ou, and Liu (2000) apply decision tree (C5.0 algorithm) and data cube technology from web log portfolios for managing classroom processes. The induction analysis discovers potential student groups that have similar characteristics and reaction to a particular pedagogical strategy. Minaei-Bidgoli and Punch (2003) classify students based on features extracted from the logged data in order to predict their final grades. They use genetic algorithms to optimize a combination of multiple classifiers by weighing feature vectors. Ueno (2004b) proposes a method of online outlier detection of learners' irregular learning processes by using the learners' response time data for the e-learning contents. The outlier detection method uses a Bayesian predictive distribution and it assists a two way instruction by using mining results for the learners' learning processes.
- *Well-known learning content management systems*. Talavera and Gaudioso (2004) propose mining student data using clustering to discover patterns reflecting user behaviors. They propose models for collaboration management to characterize similar behavior groups in unstructured collaboration spaces. Mor and Minguillon (2004) extend the sequencing capabilities of the SCORM standard to include the concept of recommended itinerary, by combining educators expertise with learned experience acquired by system usage analysis. They use clustering algorithms for grouping students. Castro, Vellido, Nebot, and Minguillon (2005) detect atypical behavior on the grouping structure of the users of a virtual campus. They propose to use a generative topographic mapping model and a clustering model to characterize groups of online students. The model neutralizes the negative impact of outliers on the data clustering process.
- *Adaptive and intelligent web-based educational systems*. Tang et al. (2000) use data clustering for web learning to promote group-based collaborative learning and to provide incremental learner diagnosis. They find clusters of students with similar learning characteristics based on the sequence and the contents of the pages they visited. Currently, they are working in smart recommendation for evolving e-learning systems (Tang & McCalla, 2005) using clustering and collaborative filtering. This is a paper recommender system that can personalize and adapt the course content based on the system's observation of the learners and the accumulated ratings given by the learners. Hamalainen et al. (2004) introduce a hybrid model, which combines both data mining and machine learning techniques in constructing a Bayesian network to describe the student's learning process. The goal is to classify students to give them differentiated guiding according to their skills and other characteristics. Beck and Woolf (2000) construct a learning agent for high-level student modeling with machine learning in intelligent tutoring systems. The agent learns to predict the probability the student's next response will be correct, and how long it will take the student to generate that response. They use linear regression to predict observable variables. Arroyo, Murray, Woolf, and Beal (2004) infer unobservable learning variables from students ITS log files. They star from a correlation analysis between variables and construct a Bayesian network that infers students' attitudes (negative and positive) and predictions of the system. They use a maximum likelihood method to learn conditional probabilities from students' data. Baker, Corbett, and Koedinger (2004) use machine-learned latent response model to detect student misuse of intelligent tutoring systems. They build a classifier to identify if a student is gaming the system in a way that leads to poor learning and in need of an intervention. Feng, Heffernan, and Koedinger (2005) look for sources of error in predicting a student's knowledge. They perform a stepwise regression to predict what metrics help to explain poor prediction of state exam scores. Muehlenbrock (2005) detects regularities and deviations in the learner's or educator's actions among others, in order to provide educators and learners with additional information to manage their learning and teaching. Damez, Marsala, Dang, and Bouchon-Meunier (2005) use a fuzzy decision tree for user modeling and discriminating a novice from an experimented user automatically. They use an agent to learn the cognitive characteristics of an user's interactions and classify users as experimented or not.

4.2.2. Association rule mining and sequential pattern mining

Association rule mining is one of the most well studied mining methods. Such rules associate one or more attributes of a dataset with another attribute, producing an if–then statement concerning attribute values. Mining association rules between sets of items in large databases was first stated by Agrawal, Imielinski, and Swami (1993) and it opened a brand new family of algorithms. The original problem was the market basket analysis that tried to find all the interesting relations between the bought products. Sequential pattern mining (Agrawal & Srikant, 1995) attempts to find inter-session patterns such as the presence of a set of items followed by another item in a time-ordered set of sessions or episodes.

These methods have been applied to web-based education systems. Associations could reveal which contents students tend to access together, or which combination of tools

they use. Sequential patterns can reveal which content has motivated the access to other contents, or how tools and contents are entwined in the learning process. Next, we describe some works about the application of these techniques in different types of web-based educational systems:

- *Particular web-based courses*. Ha et al. (2000) perform web page traversal path analysis for customized education, and web page associations for virtual knowledge structures, which can be formed by learners themselves as they navigate web pages. Yu et al. (2001) find incorrect student behavior. They modify traditional web logs, and apply fuzzy association rules to find out the relationships between each pattern of a learner's behavior; including the time spent online, number of read and published articles, number of asked questions, etc. Wang (2002) develops a portfolio analysis tool based on data mining techniques. He uses associative material clusters and sequences among them. This knowledge allows educators to study the dynamic browsing structure and to identify interesting or unexpected learning patterns. To do this, he discovers two types of relations: association relations and sequence relations between documents. Shen, Han, Yang, Yang, and Huang (2003) use data mining and case-based reasoning for distance learning. They use clustering to classify students based on their learning actions and they find sequential association rules of different knowledge points. Minaei-Bidgoli, Tan, and Punch (2004) propose mining interesting contrast rules for web-based education systems. Contrast rules help one to identify attributes characterizing patterns of performance disparity between various groups of students. Markellou, Mousourouli, Spiros, and Tsakalidis (2005) propose an ontology-based framework and discover association rules, using the a priori algorithm. The role of the ontology is to determine which learning materials are more suitable to be recommended to the user.
- *Well-known learning content management systems*. Zaïane and Luo (2001) propose the discovery of useful patterns based on restrictions, to help educators evaluate students' activities in web courses. Li and Zaïane (2004) also use recommender agents for e-learning systems which use association rule mining to discover associations between user actions and URLs. The agent recommends online learning activities or shortcuts in a course web site based on a learner's access history. Pahl and Donnellan (2003) analyze a student's individual sessions. They first define the learning period (of time) of each student and then split web server log files into individual sessions, calculate session statistics, and search for session patterns and time series. Wang, Weng, Su, and Tseng (2004) propose a four phase learning portfolio mining approach, which use sequential pattern mining, clustering and decision tree creation sequentially, to extract learning features to create a decision tree to predict which group a new learner belongs to.
- *Adaptive and intelligent web-based educational systems*. Lu (2004) uses association fuzzy rules in a personalized e-learning material recommender system. He uses fuzzy matching rules to discover associations between student's requirements and a list of learning materials. Romero et al. (2004) propose to use grammar-based genetic programming with multi-objective optimization techniques for providing a feedback to courseware authors. They discover interesting relationships from student's usage information. Merceron and Yacef (2004) use association rule and symbolic data analysis, as well as traditional SQL queries to mining student data captured from a web-based tutoring tool. Their goal is to find mistakes that often occur together. Freyberger, Heffernan, and Ruiz (2004) use association rules to guide a search for best fitting transfer model of student learning in intelligent tutoring systems. The association rules determine what operation to perform on the transfer model that predict a student's success.

4.2.3. Text mining

Text mining methods can be viewed as an extension of data mining to text data and it is very related to web content mining. It is an interdisciplinary area involving machine learning and data mining, statistics, information retrieval and natural language processing (Grobelnik, Mladenic, & Jermol, 2002). Text mining can work with unstructured or semi-structured datasets such as full-text documents, HTML files, emails, etc. Next, we describe some works on the application of these techniques in different types of web-based educational systems:

- *Particular web-based courses*. Ueno (2004a) uses data mining and text mining technologies for collaborative learning in an ILMS. She uses text mining for discussion board an expanded correspondence analysis. Learners select the relevant category which represent his/her comment and the system provides evaluations for a learner's comments between peers. Chen, Li, Wang, and Jia (2004) propose to automatically construct an e-textbook via web content mining. They use a ranking strategy to evaluate the web page suitability and they extract concept features and build concept hierarchies. Tane, Schmitz, and Stumme (2004) propose an ontology-based tool to make the most of the resources available on the web. They use text mining and text clustering techniques in order to group documents according to their topics and similarities. Hammouda and Kamel (2005) propose to perform data mining on documents, which serves as a basis for knowledge extraction in e-learning environments. In the process of text mining, a grouping (clustering) approach is also employed to identify groups of documents.
- *Well-known learning content management systems*. Dringus and Ellis (2005) propose to use text mining as a strategy for assessing asynchronous discussion forums.

Text mining techniques improve the educator's ability to evaluate the progress of a thread discussion. Bari and Benzater (2005) retrieve data from pdf interactive multimedia productions for helping the evaluation of multimedia presentations, for statistics purpose and for extracting relevant data. They identify the main blocks of multimedia presentations and retrieve their internal properties.

- *Adaptive and intelligent web-based educational systems*. Tang et al. (2000) propose to construct a personalized web tutor tree by mining both context and structure of the courseware. They use a key-word-driven text mining algorithm to select articles for distance learning students.

5. Conclusions and future research

Educational data mining is an upcoming field related to several well-established areas of research including e-learning, adaptive hypermedia, intelligent tutoring systems, web mining, data mining, etc. The application of data mining in educational systems has specific requirements not present in other domains, mainly the need to take into account pedagogical aspects of the learner and the system. Although the educational data mining is a very recent research area there is an important number of contributions published in journals, international congress, specific workshops and some ongoing books (Romero & Ventura, 2006) that show it is one new promising area. Some of the most promising work line is the use of e-learning recommendation agents (Lu, 2004; Zaïane, 2002). These recommender agents sees what a student is doing and recommends actions (activities, shortcuts, contents, etc.) they think would be beneficial to the student. Recommender agents can also be integrated in evolving e-learning systems in which materials are automatically found on the web and integrated into the system (Tang & McCalla, 2005). In this way, they help educators to detect which parts of existing materials from heterogeneous sources as the Internet are the best to use for composing new courses. Besides recommenders can also be integrated with domain knowledge and ontologies, combining web mining and semantic web in semantic web mining (Markellou et al., 2005). Semantic web mining is a successful integration of ontological knowledge at every stage of the knowledge discovery process (Becker, Vanzin, & Ruiz, 2005).

Educational data mining is a young research area and it is necessary more specialized and oriented work educational domain in order to obtain a similar application success level to other areas, such as medical data mining, mining e-commerce data, etc. We believe that some future researches lines are:

- *Mining tools more easy to use by educators or not expert users in data mining*. Data mining tools are normally designed more for power and flexibility than for simplicity. Most of the current data mining tools are too complex to use for educators and their features go well beyond the scope of what a educator may want to do. So, these tools must have a more intuitive and easy to use interface, with parameter-free data mining algorithms to simplify the configuration and execution, and with good visualization facilities to make their results meaningful to educators and e-learning designers.
- *Standardization of methods and data*. Current tools for mining data from a specific course may be useful only to its developers. There are no general tools or re-using tools or techniques that can be applied to any educational system. So, a standardization of data, and the preprocessing, discovering and postprocessing tasks is needed.
- *Integration with the e-learning system*. The data mining tool has to be integrated into the e-learning environment as another author tool. All data mining tasks (preprocessing, data mining and postprocessing) have to be carried out into a single application. Feedback and results obtained with data mining can be directly applied to the e-learning environment.
- *Specific data mining techniques*. More effective mining tools that integrate educational domain knowledge into data mining techniques. Education-specific mining techniques can help much better to improve the instructional design and pedagogical decisions. Traditional mining algorithms need to be tuned to take into account the educational context.

Acknowledgement

The authors gratefully acknowledge the financial support provided by the Spanish Department of Research of the Ministry of Science and Technology under TIN2005-08386-C05-02 Project.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data, Washington, DC* (pp. 207–216).

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Eleventh international conference on data engineering* (pp. 3–14). Taipei, Taiwan: IEEE Computer Society Press.

Arroyo, I., Murray, T., Woolf, B., & Beal, C. (2004). Inferring unobservable learning variables from students' help seeking behavior. In *Intelligent tutoring systems* (pp. 782–784).

Arruabarrena, R., Pérez, T. A., López-Cuadrado, J., & Vadillo, J. G. J. (2002). On evaluating adaptive systems for education. In *Adaptive hypermedia* (pp. 363–367).

Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., & Voyiatzaki, E. (2005). Why logging of fingertip actions is not enough for analysis of learning activities. In *Workshop on usage analysis in learning systems at the 12th international conference on artificial intelligence in education*.

Baker, R., Corbett, A., & Koedinger, K. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531–540).

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

Beck, J., & Woolf, B. (2000). High-level student modeling with machine learning. In *Intelligent tutoring systems* (pp. 584–593).

Becker, K., Ghedini, C., & Terra, E. (2000). Using kdd to analyze the impact of curriculum revisions in a Brazilian university. In *Eleventh international conference on data engineering. Proceedings of the SPIE 14th annual international conference on aerospace/defense, sensing, simulation and controls, Orlando* (pp. 412–419).

Becker, K., Vanzin, M., & Ruiz, D. D. A. (2005). Ontology-based filtering mechanisms for web usage patterns retrieval. In *6th International conference on e-commerce and web technologies* (pp. 267–277).

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education, 13*, 156–169.

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

Chen, G., Liu, C., Ou, K., & Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *Journal of Educational Computing Research, 23*(3), 305–332.

Chen, J., Li, Q., Wang, L., & Jia, W. (2004). Automatically generating an e-textbook on the web. In *International conference on advances in web-based learning* (pp. 35–42).

Damez, M., Marsala, C., Dang, T., & Bouchon-Meunier, B. (2005). Fuzzy decision tree for user modeling from human–computer interactions. In *International conference on human system learning: Who is in control?* (pp. 287–302).

Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. *Computer & Education Journal, 45*, 141–160.

Farzan, R. (2004). Adaptive socio-recommender system for open-corpus e-learning. In *Doctoral consortium of the third international conference on adaptive hypermedia and adaptive web-based systems*.

Feng, M., Heffernan, N., & Koedinger, K. (2005). Looking for sources of error in predicting student's knowledge. In *Proceedings of AAAI'05 workshop on educational data mining*.

Freyberger, J., Heffernan, N., & Ruiz, C. (2004). Using association rules to guide a search for best fitting transfer models of student learning. In *Workshop on analyzing student–tutor interactions logs to improve educational outcomes at ITS conference*.

Grob, H., Bensberg, F., & Kaderali, F. (2004). Controlling open source intermediaries – a web log mining approach. In *Proceedings of the 26th international conference on information technology interfaces* (pp. 233–242).

Grobelnik, M., Mladenic, D., & Jermol, M. (2002). Exploiting text mining in publishing and education. In *Proceedings of the ICML-2002 workshop on data mining lessons learned* (pp. 34–39).

Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In *IEEE international conference on management of innovation and technology* (pp. 715–719).

Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In *World conference on open learning and distance education, Hong Kong*.

Hammouda, K., & Kamel, M. (2005). Ch. Data mining in e-learning.

Hanna, M. (2004). Data mining in the e-learning domain. *Computers & Education Journal, 42*(3), 267–287.

Heiner, C., Beck, J., & Mostow, J. (2004). Lessons on using its data to answer educational research questions. In *Proceedings of the ITS2004 workshop on analyzing student–tutor interaction logs to improve educational outcomes* (pp. 1–9).

Heraud, J., France, L., & Mille, A. (2004). Pixed: an its that guides students with the help of learners' interaction log. In *International conference on intelligent tutoring systems (workshop analyzing student–tutor interaction logs to improve educational outcomes), Maceio* (pp. 57–64).

Hwang, W., Chang, C., & Chen, G. (2004). The relationship of learning traits, motivation and performance-learning response dynamics. *Computers & Education Journal, 42*(3), 267–287.

Iksal, S., & Choquet, C. (2005). Usage analysis driven by models in a pedagogical context.

Ingram, A. (1999). Using web server logs in evaluating instructional web sites. *Journal of Educational Technology Systems, 28*(2), 137–157.

Johnson, S., Arago, S., Shaik, N., & Palma-Rivas, N. (2000). Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments. *Journal of Interactive Learning Research, 11*(1), 29–49.

Klosgen, W., & Zytkow, J. (2002). *Handbook of data mining and knowledge discovery*. New York: Oxford University Press.

Koutri, M., Avouris, N., & Daskalaki, S. (2004). Ch. A survey on web usage mining techniques for web-based adaptive hypermedia systems.

Li, J., & Zaïane, O. (2004). Combining usage, content, and structure data to improve web site recommendation. In *International conference on e-commerce and web technologies* (pp. 305–315).

Lu, J. (2004). Personalized e-learning material recommender system. In *International conference on information technology for application* (pp. 374–379).

Luan, J. (2002). Data mining, knowledge management in higher education, potential applications. In *Workshop associate of institutional research international conference, Toronto* (pp. 1–18).

Ma, Y., Liu, B., Wong, C., Yu, P., & Lee, S. (2000). Targeting the right students using data mining. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457–464).

Markellou, P., Mousourouli, I., Spiros, S., & Tsakalidis, A. (2005). Using semantic web mining technologies for personalized e-learning experiences. In *Proceedings of the web-based education* (pp. 461–826).

Markham, S., Ceddia, J., Sheard, J., Burvill, C., Weir, J., Field, B., et al. (2003). Applying agent technology to evaluation tasks in e-learning environments. In *Proceedings of the exploring educational technologies conference*.

Mazza, R., & Milani, C. (2005). Exploring usage analysis in learning systems: Gaining insights from visualisations. In *Workshop on usage analysis in learning systems at 12th international conference on artificial intelligence in education*.

Merceron, A., & Yacef, K. (2003). A web-based tutoring tool with mining facilities to improve learning and teaching. In *Proceedings of 11th international conference on artificial intelligence in education* (pp. 201–208).

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

Merceron, A., & Yacef, K. (2005). Tada-ed for educational data mining. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 7*(1), 267–287.

Minaei-Bidgoli, B., & Punch, W. (2003). Using genetic algorithms for data mining optimization in an educational web-based system. In *GECCO* (pp. 2252–2263).

Minaei-Bidgoli, B., Tan, P., & Punch, W. (2004). Mining interesting contrast rules for a web-based educational system. In *International conference on machine learning applications*.

Monk, D. (2005). Using data mining for e-learning decision making. *Electronic Journal of e-Learning, 3*(1), 41–54.

Mor, E., & Minguillon, J. (2004). E-learning personalization based on itineraries and long-term navigational behavior. In *Proceedings of the 13th international world wide web conference* (pp. 264–265).

Mostow, J. (2004). Some useful design tactics for mining its data. In *Proceedings of the ITS2004 workshop on analyzing student–tutor interaction logs to improve educational outcomes*.

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

Muehlenbrock, M. (2005). Automatic action analysis in an interactive learning environment.

Nilakant, K., & Mitrovic, A. (2005). Application of data mining in constraint-based intelligent tutoring systems. In *Proceedings of the artificial intelligence in education, AIED* (pp. 896–898).

Pahl, C., & Donnellan, C. (2003). Data mining technology for the evaluation of web-based teaching and learning systems. In *Proceedings of the congress e-learning, Montreal, Canada*.

Paulsen, M. (2003). *Online education and learning management systems*. Bekkestua: NKI Forlaget.

Peled, A., & Rashty, D. (1999). Logging for success: Advancing the use of www logs to improve computer mediated distance learning. *Journal of Educational Computing Research, 21*(4), 413–431.

Rahkila, M., & Karjalainen, M. (1999). Evaluation of learning in computer based education using log systems. In *ASEE/IEEE frontiers in education conference, San Juan, Puerto Rico* (pp. 16–21).

Romero, C., & Ventura, S. (2006). *Data mining in e-learning*. Southampton, UK: Wit Press.

Romero, C., Ventura, S., Bra, P., & Castro, C. (2003). Discovering prediction rules in aha! courses. In *User modeling* (pp. 25–34).

Romero, C., Ventura, S., & Bra, P. D. (2004). Knowledge discovery with genetic programming for providing feedback to courseware author. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14*(5), 425–464.

Sanjeev, P., & Zytkow, J. M. (1995). Discovering enrollment knowledge in university databases. In *KDD* (pp. 246–251).

Sheard, J., Ceddia, J., Hurst, J., & Tuovinen, J. (2003). Inferring student learning behaviour from website interactions: A usage analysis. *Journal of Education and Information Technologies, 8*(3), 245–266.

Shen, R., Han, P., Yang, F., Yang, Q., & Huang, J. (2003). Data mining and case-based reasoning for distance learning. *Journal of Distance Education Technologies, 1*(3), 46–58.

Shen, R., Yang, F., & Han, P. (2002). Data analysis center based on e-learning platform. In *Proceedings of the 5th international workshop on the internet challenge: Technology and applications* (pp. 19–28).

Silva, D., & Vieira, M. (2002). Using data warehouse and data mining resources for ongoing assessment in distance learning. In *IEEE international conference on advanced learning technologies, Kazan, Russia* (pp. 40–45).

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations, 1*(2), 12–23.

Talavera, L., & Gaudioso, E. (2004). 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* (pp. 17–23).

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

Tang, C., Yin, H., Li, T., Lau, R., Li, Q., & Kilis, D. (2000). 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).

Tang, T., & McCalla, G. (2002). Student modeling for a web-based learning environment: A data mining approach. In *Eighteenth national conference on artificial intelligence, Menlo Park, CA, USA* (pp. 967–968).

Tang, T., & McCalla, G. (2005). Smart recommendation for an evolving e-learning system. *International Journal on E-Learning, 4*(1), 105–129.

Tsantis, L., & Castellani, J. (2001). Enhancing learning environments through solution-based knowledge discovery tools. *Journal of Special Education Technology, 16*(4).

Ueno, M. (2004a). Data mining and text mining technologies for collaborative learning in an ILMS "samurai". In *ICALT*.

Ueno, M. (2004b). Online outlier detection system for learning time data in e-learning and its evaluation. In *International conference on computers and advanced technology in education* (pp. 248–253).

Urbancic, T., Skrjanc, M., & Flach, P. (2002). Web-based analysis of data mining and decision support education. *AI Communications, 15*, 199–204.

Wang, F. (2002). On using data-mining technology for browsing log file analysis in asynchronous learning environment. In *Conference on educational multimedia, hypermedia and telecommunications* (pp. 2005–2006).

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

Yu, P., Own, C., & Lin, L. (2001). On learning behavior analysis of web based interactive environment. In *Proceedings of ICCEE, Oslo/Bergen, Norway*.

Zaïane, O. (2002). Building a recommender agent for e-learning systems. In *ICCE* (pp. 55–59).

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

Zaïane, O., Xin, M., & Han, J. (1998). Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in digital libraries* (pp. 19–29).

Zarzo, M. (2003). E-learning in the new era of data mining. In *International conference on network universities and e-learning, Valencia, Spain*.

Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005). Web usage mining project for improving web-based learning sites. In *Web mining workshop, Cataluna*.