Overview

Data mining in education

Cristobal Romero* and Sebastian Ventura

Applying data mining (DM) in education is an emerging interdisciplinary research field also known as educational data mining (EDM). It is concerned with developing methods for exploring the unique types of data that come from educational environments. Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes and to gain insights into and explain educational phenomena. Educational information systems can store a huge amount of potential data from multiple sources coming in different formats and at different granularity levels. Each particular educational problem has a specific objective with special characteristics that require a different treatment of the mining problem. The issues mean that traditional DM techniques cannot be applied directly to these types of data and problems. As a consequence, the knowledge discovery process has to be adapted and some specific DM techniques are needed. This paper introduces and reviews key milestones and the current state of affairs in the field of EDM, together with specific applications, tools, and future insights.

How to cite this article:
WIREs Data Mining Knowl Discov 2013, 3: 12–27 doi: 10.1002/widm.1075

INTRODUCTION

The increase of e-learning resources, instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information has created large repositories of data.[1] All this information provides a goldmine of educational data that can be explored and exploited to understand how students learn.[2] In fact, today, one of the biggest challenges that educational institutions face is the exponential growth of educational data and the use of this data to improve the quality of managerial decisions.[3]

Educational data mining (EDM) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist.[4] EDM has emerged as a research area in recent years aimed at analyzing the unique kinds of data that arise in educational settings to resolve educational research issues (Baker and Yacef, 2009). In fact, EDM, can be defined as the application of data mining (DM) techniques to this specific type of dataset that come from educational environments to address important educational questions.[5,6]

EDM analyze data generated by any type of information system supporting learning or education (in schools, colleges, universities, and other academic or professional learning institutions providing traditional and modern forms and methods of teaching, as well as informal learning). These data[7] are not restricted to interactions of individual students with an educational system (e.g., navigation behavior, input in quizzes and interactive exercises) but might also include data from collaborating students (e.g., text chat), administrative data (e.g., school, school district, teacher), demographic data (e.g., gender, age, school grades), student affectivity (e.g., motivation, emotional states), and so forth. These data have typical characteristics such as multiple levels of hierarchy (subject, assignment, question levels), context (a particular student in a particular class encountering a particular question at a particular time on a particular date), fine grained (recording of data at different resolutions to facilitate different analyses, e.g., recording data every 20 second), and longitudinal (much data recorded over many sessions for a long period of time, e.g., spanning semester and year-long courses).

EDM is an interdisciplinary area including but not limited to information retrieval, recommender systems, visual data analytics, domain-driven DM, social network analysis (SNA), psychopedagogy,

*Correspondence to: cromero@uco.es
Department of Computers Science and Numerical Analysis, University of Cordoba, Cordoba, Spain.

DOI: 10.1002/widm.1075

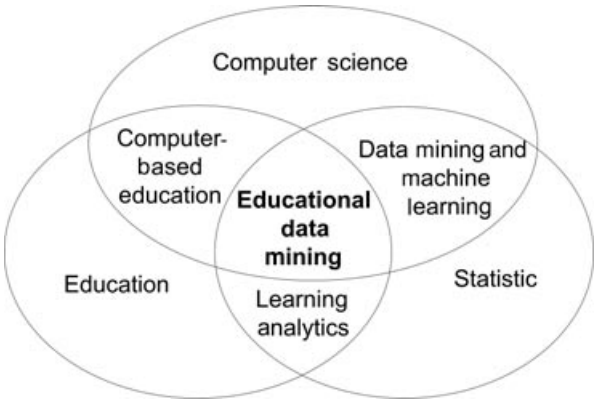

FIGURE 1 | Main areas related to educational data mining.

cognitive psychology, psychometrics, and so on. In fact, EDM can be drawn as the combination of three main areas (see Figure 1): computer science, education, and statistics. The intersection of these three areas also forms other subareas closely related to EDM such as computer-based education, DM and machine learning, and learning analytics (LA).

Of all the aforementioned areas (see Figure 1), the field most related to EDM is LA, also known as academic analytics.[8] LA is focused on data-driven decision-making and integrating the technical and the social/pedagogical dimensions of LA.[9] However, although EDM is generally looking for new patterns in data and developing new algorithms and/or models, LA is applying known predictive models in instructional systems.[10] In fact, LA can be defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. Although LA and EDM can share many attributes and have some similar goals and interests, the next key differences can be distinguished between both communities[9]:

- *Techniques*: In LA, the most used techniques are statistics, visualization, SNA, sentiment analysis, influence analytics, discourse analysis, concept analysis, and sense-making models. In EDM, the most used techniques are classification, clustering, Bayesian modeling, relationship mining and discovery with models.
- *Origins*: LA has stronger origins in Semantic Web, intelligent curriculum, and systemic interventions. EDM has strong origins in educational software, student modeling, and predicting course outcomes.
- *Emphasis*: LA has more emphasis on the description of data and results; however, EDM has more emphasis on the description and comparison of the DM techniques used.
- *Type of discovery*: In LA, leveraging human judgment is key; automated discovery is a tool used to accomplish this goal. In EDM, automated discovery is key; leveraging human judgment is a tool used to accomplish this goal.

This paper provides an updated overview of the current state of knowledge in EDM with the objective of introducing it to researchers, instructors and advanced students without a strong background in the field. The paper is organized as follows. First, the background of EDM is described. Then the main types of educational environments and their data are shown. The following sections describe the main goals and the specific knowledge discovery process in EDM. Next, the most popular methods used in EDM are presented. Subsequently, some examples of applications or tasks in educational environments and some examples of specific DM tools are listed. Finally, some future lines of research and conclusions are outlined.

BACKGROUND

EDM has emerged as an independent research area in recent years, starting with research in intelligent tutoring systems (ITS), artificial intelligence in education (AIED), user modeling (UM), technology-enhanced learning (TEL), and adaptive and intelligent educational hypermedia (AIEH). Its origins lie in a series of workshops (see Table 1) organized into related conferences that began in 2000. The first workshop, referred to as 'Educational Data Mining', took place in 2005 and culminated in 2008 with the establishment of the annual International Conference on Educational Data Mining organized by the International Working Group on Educational Data Mining.

The first conference EDM2008 was held in Montreal, Canada; then EDM2009 in Cordoba, Spain; EDM2010 in Pittsburgh, USA; EDM2011 in Eindhoven, The Netherlands; EDM2012 in Chania, Greece; and the next EDM2013 will be held in Memphis EEUU. There are some other closely related conferences (see Table 2) in which EDM is colocated most years. All of them are older than EDM with the exception of the LAK conference (International Conference on Learning Analytics and Knowledge),

TABLE 1 | Educational Data Mining (EDM) Workshops

Title	Acronym	Location	Year
Workshop on Applying Data Mining in e-Learning	EC-TEL'07-ADML	Crete, Greece	2007
Workshop on Educational Data Mining	AIED'07-EDM	California, USA	2007
Workshop on Educational Data Mining	ICALT'07-EDM	Niigata, Japan	2007
Workshop on Educational Data Mining	AAAI'05-EDM	Boston, USA	2006
Workshop on Educational Data Mining	ITS'06-EDM	Jhongli, Taiwan	2006
Workshop on Educational Data Mining	AAAI'05-EDM	Pittsburgh, USA	2005
Workshop on Usage Analysis in Learning Systems	AIED'05-W1	Amsterdam, the Netherlands	2005
Workshop on Analyzing Student–Tutor Interaction Logs to Improve Educational	ITS'04-W2	Maceio, Brazil	2004
Workshop on Applying Machine Learning to ITS Design/Construction	ITS'00-W3	Montreal, Canada	2000

TABLE 2 | Related Conferences about Educational Data Mining

Title	Acronym	Type	Year
International Conference on Educational Data Mining	EDM	Annual	2008
International Conference on Learning Analytics and Knowledge	LAK	Annual	2011
International Conference on Artificial Intelligence in Education	AIED	Biannual	1982
International Conference on Intelligent Tutoring Systems	ITS	Biannual	1988
International Conference on User Modeling, Adaptation, and Personalization	UMAP	Annual	2009

which is younger. The first LAK conference, was in Banff, Canada, in 2011 and the second in Vancouver, Canada, in 2012.

Currently, only two books on EDM have been published. The first, entitled *Data Mining in E-Learning*,[11] has 17 chapters oriented to Web-based educational environments. The second, entitled *Handbook of Educational Data Mining* (Romero et al., 2010), and has 36 chapters oriented to different types of educational settings.

There are several surveys in journals and chapters in books about EDM. The first and most popular review of EDM research was presented in a journal,[5] and was followed by a more theoretical paper (Baker and Yacef, 2009) and a more complete review.[6] A first and wide-ranging book chapter review was oriented to the application of DM in e-learning,[12] a second and shorter book chapter was more oriented to ITS[13] and a third book chapter was the most generic but the shortest.[7] Finally, a recent report was published by the US Office of Educational Technology about how to enhance teaching and learning through EDM and LA.[10]

There are a wide range of international and prestigious journals in which a large number of EDM papers have been published (see Table 3). Of all of them, the most specific is the Journal of Educational Data Mining (http://www.educationaldatamining.org/JEDM/), which was launched in 2009 as an online and free journal. On the other hand, a selection of the most cited papers in EDM area is shown in Table 4.

There are an increasing number of important authors in EDM area as well as the authors of this paper. Ryan Baker from Worcester Polytechnic Institute, USA, that is, the president of the EDM society. Kalina Yacef from the University of Sydney, Australia, that is, the editor of JEDM journal and member of the steering committee of EDM society together with Tiffany Barnes from University of North Calorina, USA; Joseph E. Beck from Worcester Polytechnic Institute, USA; Michel Desmarais from Ecole Polytechnic de Montreal, Canada; Neil Heffernan from Worcester Polytechnic Institute, USA; Agathe Merceron from Beuth University of Applied Sciences, Germany; and Mykola Pechenizkiy from Eindhoven University of Technology, the Netherlands. Other relevant authors are Osmar Zaiäine from Alberta University, Canada; John Stamper from Carnegie Mellon University, USA; Judy Kay from The

TABLE 3 | Some examples of Educational Data Mining Related Journals

Journal Title	Acronym	Editorial	Impact Factor 2011
Journal of Educational Data Mining	JEDM	EDM Society	–
Journal of Artificial Intelligence in Education	JAIED	AIED Society	–
Journal of the Learning Sciences	JLS	Taylor&Francis	2.000[1]
Computer and Education	CAE	Elsevier	2.621[2]
IEEE Transactions on Learning Technologies	TLT	IEEE	–
IEEE Transactions on Knowledge and Data Engineering	KDE	IEEE	1.657[2]
ACM Special Interest Group on Knowledge Discovery and Data Mining, Explorations	SIGKDD Explorations	ACM	–
User Modeling and User-Adapted Interaction	UMUAI	Springer	1.400[2]
Internet and Higher Education	INTHIG	Elsevier	1.015[1]
Decision Support Systems	DCS	Elsevier	1.687[2]
Expert Systems with Applications	ESWA	Elsevier	2.203[2]
Knowledge-Based Systems	KBS	Elsevier	2.422[2]

[1] JCR Social Science Edition.
[2] JCR Science Edition.

TABLE 4 | Top 10 Most Cited Papers about Educational Data Mining until August 2012

Paper Title	Reference	Number of citations[1]	Number of citations[2]
Educational data mining: a survey from 1995 to 2005	5	296	158
Data mining in course management systems: Moodle case study and tutorial	14	191	83
Web usage mining for a better Web-based learning environment	15	183	–
Off-task behavior in the cognitive tutor classroom: when students game the system	16	177	25
Building a recommender agent for e-learning systems	17	168	–
Detecting student misuse of intelligent tutoring systems	18	156	20
The ecological approach to the design of e-learning environments: purpose-based capture and use of information about learners	19	136	–
Student modeling and machine learning	20	127	–
Towards evaluating learners' behavior in a Web-based distance learning environment	21	117	–
Smart recommendation for an evolving e-learning system: architecture and experiment	22	98	–

[1] Google Schoolar.
[2] SciVerse Scopus.

University of Sydney, Australia; Kenneth Koedinger and Jack Mostow from Carnegie Mellon University, USA; Rafi Nachmias from Tel Aviv University, Israel; Gord McCalla from University of Saskatchewan, Canada; Arthur Graesser from The University of Hemphis, USA; and so forth.

There are several international societies related to EDM. The most important are the International Educational Data Mining Society (http://www.educationaldatamining.org) which was founded by the International Working Group on Educational Data Mining in 2011; the Society for Learning Analytics Research (SoLAR) (http://www.solaresearch.org/) which was created in 2011; and the IEEE Task Force of Educational Data Mining (EDM-TF) (http://datamining.it.uts.edu.au/edd/) which was created in 2012.

Finally, to demonstrate the current increasing interest in EDM, Figure 2 shows the number of references or results that return a freely accessible Web search engine such as *Google Schoolar* and a subscription-based tool such as *SciVerse Scopus* when searching the exact term 'Educational Data Mining' in each year from 2004 to 2011. As can be seen, both numbers grow in an exponential way, showing the high interest in this topic, and in the last two years the number of cites in *SciVerse Scopus* is higher than in *Google Schoolar*.

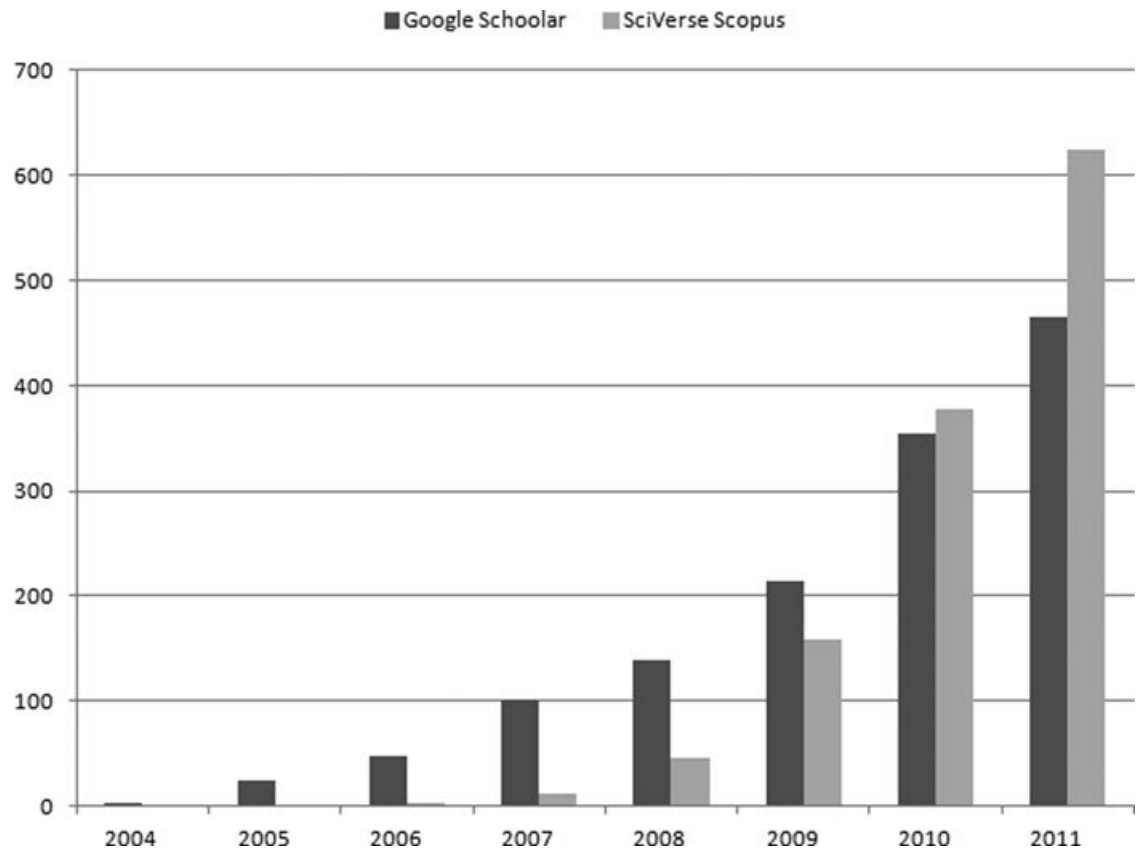

FIGURE 2 | Number of educational data mining references in Google Schoolar and cites in SciVerse Scopus by year.

BOX 1: EDM COMPETITIONS

There were two specific international EDM competitions with the same objective of predicting whether a student will answer the next test question correctly. The first competition was the KDD Cup 2010 (https://pslcdatashop.web.cmu.edu/KDDCup/) on the Educational Data Mining Challenge with 5096 participants. The data comes from 10,000 students of Carnegie Learning Inc.'s Cognitive Tutors. And the second competition was the Kaggle Competition (http://www.kaggle.com/c/WhatDoYouKnow) with 252 teams and a prize pool of $5000. The data in this competition comes from students studying for three tests: the GMAT, SAT, and ACT.

TYPES OF EDUCATIONAL ENVIRONMENTS

Nowadays, there is a wide variety of educational environments and information systems both in traditional education and computer-based education (see Figure 3). Each one of them provides different data sources that have to be pre-processed in different ways depending on both the nature of available data and the specific problems and tasks to be resolved by DM techniques.[5]

Traditional Education

Traditional education or back-to-basics refers to long-established customs found in schools that society has traditionally deemed to be appropriate. These environments are the most widely used educational system, based mainly on face-to-face contact between educators and students organized through lectures, class discussion, small groups, individual seat work, and so forth. These systems gather information on student attendance, marks, curriculum goals, and individualized plan data. Also, educational institutions store many diverse and varied sources of information[23]: administrative data in traditional databases (with a student's information, the educator's information, class and schedule information, etc.), online information (online Web pages and course content pages), and so forth. In conventional classrooms, educators normally attempt to enhance instruction by monitoring students' learning processes and analyzing their performance on paper and through observation.[24]

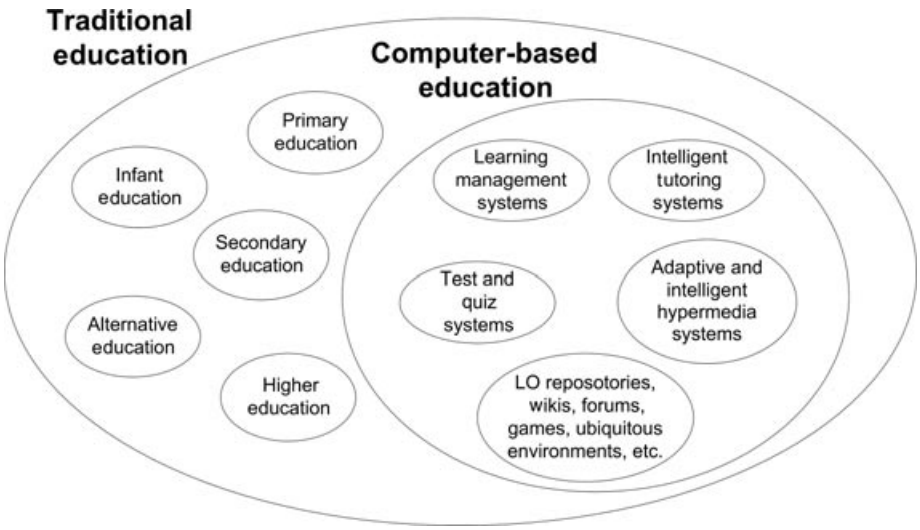

FIGURE 3 | Types of traditional and computer-based educational environments and systems.

TABLE 5 | Types of Traditional Educational Environments

System	Description
Infant/preschool education	This provides learning to children before statutory and obligatory education, usually between the ages of zero and three or five.
Primary/elementary education	This consists of the first 5–7 years of formal, structured education.
Secondary education	This comprises the formal education that occurs during adolescence.
Higher/tertiary education	This is the noncompulsory educational level that follows the completion of a school providing a secondary education.
Alternative/special education	This includes not only forms of education designed for students with special needs, but also forms of education designed for a general audience and employing alternative educational methods.

Some examples of traditional education systems are described in Table 5. Finally, it is important to note that all these traditional systems can also use computer-based educational systems as a complementary tool to face-to-face sessions.

Computer-Based Educational Systems

Computer-based education (CBE) means using computers in education to provide direction, to instruct or to manage instructions given to the student. CBE systems were originally stand-alone educational applications that ran on a local computer without using artificial intelligence techniques for student modeling, adaptation, personalization, and so forth. On the one hand, the global use of Internet has led to today's plethora of new Web-based educational systems such as e-learning systems, e-training systems, online instruction systems, and so forth. On the other hand, the increasing use of artificial intelligence techniques has induced the emergence of new intelligent and adaptive educational systems. Some of the main types of computer-based educational systems used currently are (see Table 6 for a description) learning and management systems (LMS),[14] ITS,[2] adaptive and intelligent hypermedia systems (AIHS),[25] test and quiz systems,[26] and other types of CBE systems.

GOALS

Data mining has already been successfully applied to other areas or domains such as business, bioinformatics, genetics, medicine, and so forth. Although the discovery methods used in all these areas can be seen similar, the objectives are different.[6] For instance, in comparing the use of data mining in e-commerce versus EDM. The main objective of data mining in e-commerce is to increase profit. Profit is a tangible goal that can be measured in terms of sums of money,

TABLE 6 | Types of Computer-Based Educational Systems

System	Description
Learning management systems	Suites of software that provide course-delivery functions: administration, documentation, tracking, and reporting of training programs—classroom and online events, e-learning programs, and training content. They also offer a wide variety of channels and workspaces to facilitate information sharing and communication among all the participants in a course. They record any student activities involved, such as reading, writing, taking tests, performing tasks in real, and commenting on events with peers.
Intelligent tutoring systems (ITS)	ITS provide direct customized instruction or feedback to students by modeling student behavior and changing its mode of interaction with each student based on its individual model. Normally, it consists of a domain model, student model, and pedagogical model. ITS record all student–tutor interaction (mouse clicks, typing, and speech).
Adaptive and intelligent hypermedia (AIH) systems	These attempt to be more adaptive by building a model of the goals, preferences, and knowledge of each individual student and using this model throughout interaction with the student to adapt to the needs of that student. The data recorded by AIHs are similar to ITS data.
Test and quiz systems	The main goal of these systems is to measure the students' level of knowledge with respect to one or more concepts or subjects by using a series of questions/items and other prompts for the purpose of gathering information from respondents. They store a great deal of information about students' answers, calculated scores, and statistics.
Other types	Learning object repositories, concept maps, social networks, wikis, forums, educational games, virtual reality/3D, ubiquitous computing, and so forth.

TABLE 7 | Example of Users/Stakeholders and Objectives

User/Stakeholders	Examples of objectives
Learners	To support a learner's reflections on the situation, to provide adaptive feedback or recommendations to learners, to respond to student needs, to improve learning performance, and so on
Educators	To understand their students' learning processes and reflect on their own teaching methods, to improve teaching performance, to understand social, cognitive and behavioral aspects, and so on
Researchers	To develop and compare data mining techniques to be able to recommend the most useful one for each specific educational task or problem, to evaluate learning effectiveness when using different settings and methods, and so on.
Administrators	To evaluate the best way to organize institutional resources (human and material) and their educational offer, and so on

and which leads to clear secondary measures such as the number of customers and customer loyalty. As the main objective of data mining in education is largely to improve learning, measurements are more difficult to obtain, and must be estimated through proxies such as improved performance. So, in general it enables data-driven decision-making for improving current educational practice and learning materials. However, there are many more specific objectives in EDM depending on the viewpoint of the final user and the problem to resolve. Some examples of particular problems are[27]

- How to (re)organize classes, or assessment, or the placement of materials based on usage and performance data.
- How to identify those who would benefit from feedback, study advice, or other help provided.
- How to decide which kind of help, feedback, or advice would be most effective.
- How to help learners in finding and searching useful material, individually, or in collaboration with peers.

Although an initial consideration seems to involve only two main groups of potential users/stakeholders—the learners and the instructors—there are actually more groups involved with many more objectives, as can be seen in Table 7.

The number of possible problems or objectives for each type of stakeholder is huge. For example,

TABLE 8 | Current Topics of Interest of Educational Data Mining Research Community

Topics of Interest	Description
Generic frameworks and methods	To develop tools, frameworks, methods, algorithms, approaches, and so forth, specifically oriented to educational data mining research.
Mining educational data	Mining assessment data, mining browsing or interaction data, mining the results of educational research (e.g., A/B tests), and so forth.
Educational process mining	To extract process-related knowledge from event logs recorded by educational systems.
Data-driven adaptation and personalization	To apply data mining methods and techniques for improving adaptation and personalization in educational environments and systems.
Improving educational software	Many large educational data sets are generated by computer software. Can we use our discoveries to improve the software's effectiveness?
Evaluating teaching interventions	Student learning data provides a powerful mechanism for determining which teaching actions are successful. How can we best use such data?
Emotion, affect, and choice	The student's level of interest is critical. Can we detect when students are bored and uninterested? What other affective states or student choices should we track?
Integrating data mining and pedagogical theory	Data mining typically involves searching a large volume of models. Can we use existing educational and psychological knowledge to better focus our research?
Improving teacher support	What types of assessment information would help teachers? What types of instructional suggestions are both feasible to generate and would be welcomed by teachers?
Replication studies	To apply a previously used technique to a new domain, or to reanalyze an existing data set with a new technique.
Best practices	Best practices for adaptation of data mining, information retrieval, recommender system, opinion mining, and question answering techniques to educational context.

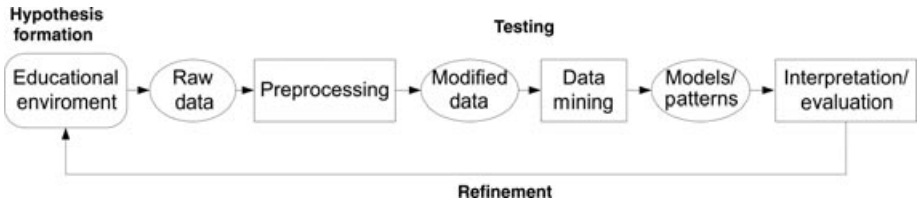

FIGURE 4 | Educational knowledge discovery and data mining process.

from the point of view of EDM researchers, there is a wide range of current topics of interest (see Table 8).

EDUCATIONAL KNOWLEDGE DISCOVERY PROCESS

The process of applying data mining to educational systems can be interpreted from different points of view (Romero et al., 2010).

On the one hand, from an educational and an experimental viewpoint, it can be seen as an iterative cycle of hypothesis formation, testing, and refinement (see Figure 4). In this process, the goal is not just to turn data into knowledge, but also to filter mined knowledge for decision-making about how to modify the educational environment to improve student's learning. This is a type of formative evaluation of an educational program while it is still in development, and with the purpose of continually improving the program. Analyzing how students use the system is one way to evaluate instructional design in a formative manner and may help educational designers to improve instructional materials. For example, EDM techniques discover models/patterns that can be used to assist educational designers to establish a pedagogical basis for decisions when designing or modifying an environment's pedagogical approach.

On the other hand, from a DM viewpoint, it can be seen very similar to the general knowledge discovery and data mining (KDD) process (see Figure 4) although there are important differences or specific characteristics in each step as is described in the following subsections.

Educational Environment

Depending on the type of the educational environment (traditional classroom education, computer-based or Web-based education) and an information

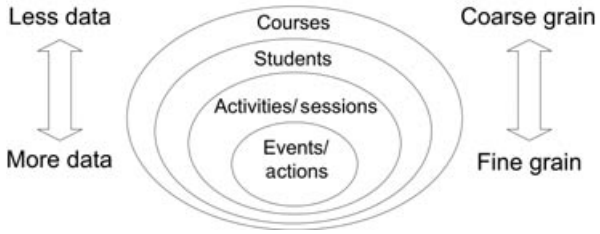

FIGURE 5 | Different levels of granularity and their relationship to the amount of data.

system that supports it (a learning management, intelligent tutoring or adaptive hypermedia system) different kinds of data can be collected to resolve different educational problems.[4] All these data may come from different sources including administrative data, field observations, motivational questionnaires, measurements collected from controlled experiments, final marks, and so on. Gathering and integrating this raw data for mining are nontrivial tasks on their own and thus a preprocessing step is necessary.

Preprocessing

In educational contexts, it is natural for data preprocessing to be a very important and complicated task, and sometimes the data preprocessing itself takes up more than half of the total time spent solving the data mining problem.[10] First, the educational data available (raw, original, or primary data) to solve a problem is not in the appropriate form (or abstraction). And second, given the heterogeneous and hierarchical nature of educational data, determining data structures and formats that represent an event under consideration become key and the best data structure will also depend on the type of problem to be solved. So, it is necessary to convert the data to an appropriate form (modified data) for solving a specific educational problem. This includes choosing what data to collect, focusing on the questions to be answered, and making sure the data align with the questions. On the other hand, educational environments can store a huge amount of potential data from multiple sources with different formats and with different granularity levels (from coarse to fine grain) or multiple levels of meaningful hierarchy (keystroke level, answer level, session level, student level, classroom level, and school level) that provide more or less data (see Figure 5). So, it can be necessary to carry out data integration at the appropriate granularity level. Normally, also available are a huge number of variables/attributes with information about each student, which can be reduced into a summary table for better analysis. Continuous attributes are normally transformed/discretized into categorical attributes to improve their comprehensibility. Issues of time, sequence, and context also play important roles in the study of educational data. Time is important to capture data such as length of practice sessions or time to learn. Sequences represent how concepts build on one another and how practice and tutoring should be ordered. Context is important for explaining results and knowing where a model may or may not work. Finally, it is important to maintain and protect the confidentiality of student information when integrating all collected data by deleting some personal information (not useful for mining purposes) such as name, e-mail, telephone number, and so on, and thus anonymizing data by using, for example, a numerical sequence for identifying students.

Data Mining

The majority of traditional data mining techniques including but not limited to classification, clustering, and association analysis techniques have been already applied successfully in the educational domain.[13] Nevertheless, educational systems have special characteristics that require a different treatment of the mining problem. For example, methods for hierarchical data mining and longitudinal data modeling have to be used in EDM. As a consequence, some specific data mining techniques are needed to address learning and other data about learners. However, EDM is still an emerging research area, and we can foresee that its further development will result in a better understanding of the challenges specific to this field and will help researchers involved in EDM to see which techniques can be adopted and what new customized techniques have to be developed. On the other hand, there are some data mining methods that are more appropriate for solving some of the types of educational problems to resolve, as described in *Methods*.

Interpretation of Results

This final step is very important to apply the knowledge acquired to making decision about how to improve the educational environment or system.[28] So the models obtained by the DM algorithms have to be comprehensible and useful for the decision-making process. For example, white-box DM models such as decision trees are preferable to black-box models such as neural networks as they are more accurate but less comprehensible. Visualization techniques are also very useful for showing results in a way that is easier to interpret. For example, it is better to show only a subset of association rules in graphic format instead of showing all the rules discovered (normally hundreds or thousands) in a traditional text format.

Finally, recommender systems can be the best way to display results, information, explanations, recommendations and comments to a nonexpert user in DM such as instructors. Thus instead of showing the obtained DM model, a list of suggestions or conclusions about the results and how to apply them are shown to the users.

METHODS

There are a number of popular methods within EDM.[5,7,13,29] Some of them are widely acknowledged to be universal across types of data mining, such as prediction, clustering, outlier detecting, relationship mining, SNA, process mining, and text mining. And others have particular prominence within EDM, such as the distillation of data for human judgment, discovery with models, knowledge tracing (KT) and nonnegative matrix factorization.

Prediction

The goal of prediction is to infer a target attribute or single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). Types of predictions methods are classification (when the predicted variable is a categorical value), regression (when the predicted variable is a continuous value), or density estimation (when the predicted value is a probability density function). In EDM, prediction has been used for forecasting student performance[30] and for detecting student behaviors.[31]

Clustering

The goal of clustering is to identify groups of instances that are similar in some respect. Typically, some kind of distance measure is used to decide how similar instances are. Once a set of clusters has been determined, new instances can be classified by determining the closest cluster. In EDM, clustering can be used for grouping similar course materials or grouping students based on their learning and interaction patterns.[32]

Outlier Detection

The goal of outlier detection is to discover data points that are significantly different than the rest of data. An outlier is a different observation (or measurement) that is usually larger or smaller than the other values in data. In EDM, outlier detection can be used to detect students with learning difficulties, deviations in the learner's or educator's actions or behaviors, and for detecting irregular learning processes.[33]

Relationship Mining

The goal of relationship mining is to identify relationships between variables and normally to encode them in rules for later use. There are different types of relationship in mining techniques such as association rule mining (any relationship between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationship between variables). In EDM, relationship mining has been used to identify relationships in learners' behavior patterns and diagnosing students' learning difficulties or mistakes that frequently occur together.[34]

Social Network Analysis

The goal of SNA is to understand and measure the relationships between entities in networked information. SNA views social relationships in terms of network theory consisting of nodes (representing individual actors within the network) and connections or links (which represent relationships between the individuals, such as friendship, kinship, organizational position, sexual relationships, etc.). In EDM, SNA can be used for mining to interpret and analyze the structure and relations in collaborative tasks and interactions with communication tools.[35]

Process Mining

The goal of process mining is to extract process-related knowledge from event logs recorded by an information system to have a clear visual representation of the whole process. It consists of three subfields: conformance checking, model discovery, and model extension. In EDM, process mining can be used for reflecting students behavior in terms of their examination traces consisting of a sequence of course, grade, and timestamp triplets for each student.[36]

Text Mining

The goal of text mining, also referred to as text data mining or text analytics, is to derive high-quality information from text. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling. In EDM, text mining has been used to analyze the content of discussion boards, forums, chats, Web pages, documents, and so forth.[37]

Distillation of Data for Human Judgment

The goal is to represent data in intelligible ways using summarization, visualization and interactive interfaces to highlight useful information and support

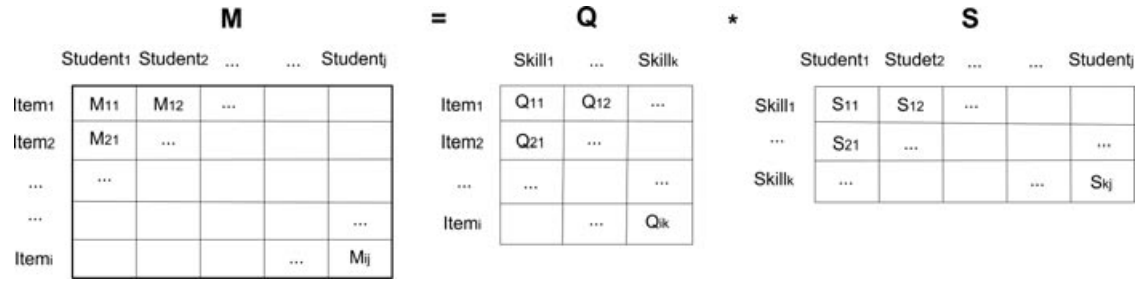

FIGURE 6 | Example of nonnegative matrix factorization and *Q*-matrix interpretation.

decision-making. On the one hand, it is relatively easy to obtain descriptive statistics from educational data to obtain global data characteristics and summaries and reports on learner behavior. On the other hand, information visualization and graphic techniques help to see, explore, and understand large amounts of educational data at once. In EDM, it also known as distillation for human judgment[13] and it has been used for helping educators to visualize and analyze the students' course activities and usage information.[38]

Discovery with Models

The goal of discovering with models is to use a previously validated model of a phenomenon (using prediction, clustering, or manual knowledge engineering) as a component in another analysis such as prediction or relationship mining.[29] It is particularly prominent in EDM and it supports the identification of relationships between student behaviors and students' characteristics or contextual variables, the analysis of research questions across a wide variety of contexts, and the integration of psychometric modeling frameworks into machine-learning models.[10]

Knowledge Tracing

KT is a popular method for estimating student mastery of skills that has been used in effective cognitive tutor systems.[39] It uses both a cognitive model that maps a problem-solving item to the skills required, and logs of students' correct and incorrect answers as evidence of their knowledge on a particular skill. KT tracks student knowledge over time and it is parameterized by four variables. There is an equivalent formulation of KT as a Bayesian network.

Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a technique that allows a straightforward interpretation in terms of a Q-Matrix, also termed a transfer model.[40] There are many NMF algorithms and they can yield different solutions. NMF consists of a matrix of positive numbers, as the product of two smaller matrices. For example, in the context of education, a matrix **M** that represents the observed examinee's test outcome data (see Figure 6) that can be decomposed into two matrices: **Q** that represents the *Q*-matrix of items and **S** that represents each student's mastery of skills.

APPLICATIONS

There are many examples of applications or tasks in educational environments that can be resolved through DM.[3,4] Between all of them, predicting students' performance is the oldest and most popular application of DM in education. However, in recent years, EDM has been applied to address a large number of new and different problems (see Table 9).

SPECIFIC DATA MINING TOOLS

Nowadays, there are a lot of general free and commercial DM tools and frameworks[49] that can be used for mining datasets from any domain or research area. However, all these tools are not specifically designed for pedagogical/educational purposes and problems. So they are cumbersome for an educator to use because they are designed more for power and flexibility than for simplicity. However, an increasing number of mining tools have been developed that are specifically oriented to solve different educational problems (see Table 10).

CONCLUSIONS AND FUTURE INSIGHTS

EDM brings together an interdisciplinary community of computer scientists, learning scientists, psychometricians, and researchers from other fields. EDM applies techniques coming from statistics, machine learning, and data mining to analyze data collected during teaching and learning, tests learning theories, and informs decision-making in educational practice. The field of EDM has grown substantially in recent years with two related annual conferences, two books, several surveys in books and journals, two

TABLE 9 | Some Examples of Educational Data Mining Tasks or Applications

Task/Application	Description/Goal	Reference
Predicting student performance	To estimate the unknown value of a student's performance, knowledge, score or mark	26
Scientific inquiry	To develop and test scientific theories on technology-enhanced learning, to formulate new scientific hypotheses, and so on	9
Providing feedback for supporting instructors	To provide feedback to support educators in decision-making about how to improve students' learning and enable them to take appropriate proactive and/or remedial action	26
Personalizing to students	To adapt automatically learning, navigation, content, presentation, and so forth, to each particular students	41
Recommending to students	To make recommendations to students with respect to their activities or tasks, links to visits, problems or courses to be done, and so forth.	42
Creating alerts for stakeholders	To monitor students' learning progress for detecting in real time undesirable student behaviors such as low motivation, playing games, misuse, cheating, dropping out, and so forth	43
User/Student modeling	To develop and tune cognitive models of human students that represent their skills and declarative knowledge	44
Domain modeling	To describe the domain of instruction in terms of concepts, skills, learning items and their interrelationships	45
Grouping/Profiling students	To create groups of students according to their customized features, personal characteristics, personal learning data, and so forth	46
Constructing courseware	To help instructors and developers to carry out the construction/development process of courseware and learning content automatically	28
Planning and scheduling	To plan future courses, student course scheduling, planning resource allocation, admission and counseling processes, developing curriculum, and so forth	47
Parameter estimation	To infer parameters of probabilistic models from given data to predict the probability of events of interest	48

TABLE 10 | Examples of Educational Data Mining Tools

Tool	Goal	Reference
EPRules	To discover prediction rules to provide feedback for courseware authors	50
GISMO	To visualize what is happening in distance learning classes	38
TADA-ED	To help teachers to identify relevant patterns in students' online exercises	51
O3R	To retrieve and interpret sequential navigation patterns	52
Synergo/ColAT	To analyze and produce interpretative views of learning activities	53
LISTEN Mining tool	To explore large student–tutor interaction logs	54
MINEL	To analyze navigational behavior and the performance of the learner	55
LOCO-Analyst	To provide teachers with feedback on the learning process	56
Measuring tool	To measure the motivation of online learners	57
DataShop	To store and analyze click-stream data, fine-grained longitudinal data generated by educational systems	1
Decisional tool	To discover factors contributing to students' success and failure rates	58
CIECoF	To make recommendations to courseware authors about how to improve courses	28
SAMOS	To browse student activity using overview spreadsheets	59
PDinamet	To support teachers in collaborative student modeling	60
AHA! Mining Tool	To recommend the best links for a student to visit next	61
EDM Visualization Tool	To visualize the process in which students solve procedural problems in logic	62
Meerkat-ED	To analyze participation of students in discussion forums using social network analysis techniques	35
MMT tool	To facilitate the execution of all the steps in the data mining process of Moodle data for newcomers	63
SNAPP	To visualize the evolution of participant relationships within discussions forums	64
AAT	To access and analyze students' behavior data in learning systems	65
DRAL	To discover relevant e-activities for learners	66
E-learning Web Miner	To discover student's behavior profiles and models about how they work in virtual courses	67

competitions (see Box 1), an increasing number of specific tools, and so forth. Time will tell if this evolution continues or not, and if other communities such as KDD perceive it as yet another application domain of data mining or really as a new subfield. Currently, there is a large amount of work to be done in the EDM community in order for it to be considered as a mature area. EDM has to be much more widely used and applied, not only by researchers but also by teachers and institutions. Although EDM has been used in some courses and institutions with success, it is necessary to move from the lab to the general market, and to achieve this objective it is necessary to carry out the next stages of future work:

On the one hand, it is essential that EDM tools are open source or freely available to download in order for them to be used by a much wider and broader population. In fact, most of the current specific EDM tools (see Table 10) are not available for download. EDM tools must be included and integrated into their own computer-based educational systems alongside another tools such as course designer tools, test generator tools, report tools, and so forth. EDM tools must also be easier for educators to use. Usually, they are required to select the specific DM method/algorithm they want to apply/use. And these DM algorithms usually require parameters and they have to provide appropriate values in advance to obtain good results/models. So, the educators must possess a certain amount of expertise to find the right settings. A solution to this problem is the use of decision support systems, wizard tools, recommendation engines and free-parameter DM algorithms to automate and facilitate all the EDM processes for instructors.

On the other hand, educators and institutions should develop a data-driven culture of using data for making instructional decisions and improving instruction. Results from EDM research are typically achieved in the narrow context of specific research projects and educational settings. However, it is necessary to obtain more general results, for instance, whether the same student model parameters also can be used with other student populations, or whether a predictive model is still reliable when used in a different context. There is therefore an increasing need for replication studies to test for broader generalizations. As a practical consequence of this need, EDM researchers have become increasingly more interested in open data repositories and standard data formats to promote the exchange of data and models.

ACKNOWLEDGMENTS

This research is supported by projects of the Regional Government of Andalucia and the Ministry of Science and Technology, P08-TIC-3720 and TIN-2011-22408, respectively, and FEDER funds.

REFERENCES

1. Koedinger K, Cunningham K, Skogsholm A, Leber B. An open repository and analysis tools for fine-grained, longitudinal learner data. In: *First International Conference on Educational Data Mining*. Montreal, Canada; 2008, 157–166.
2. Mostow J, Beck J. Some useful tactics to modify, map and mine data from intelligent tutors. *J Nat Lang Eng* 2006, 12:195–208.
3. Bala M, Ojha DB. Study of applications of data mining techniques in education. *International J Res Sci Technol* 2012, 1: 1–10.
4. Romero C, Ventura S, Pechenizky M, Baker R. *Handbook of Educational Data Mining*. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press; 2010.
5. Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. *J Expert Syst Appl* 2007, 1:135–146.
6. Romero C, Ventura S. Educational data mining: a review of the state-of-the-art. *IEEE Trans Syst Man Cybern C: Appl Rev* 2010, 40:601–618.
7. Scheuer O, McLaren BM. Educational data mining. In: *The Encyclopedia of the Sciences of Learning*. New York, NY: Springer; 2011.
8. Baepler P, Murdoch CJ. Academic analytics and data mining in higher education. *Int J Schoolarship Teach Learn* 2010, 4:1–9.
9. Siemens G, Baker RSJd. Learning analytics and educational data mining: towards communication and collaboration. In: *Proceedings of the 2nd International*

Conference on Learning Analytics and Knowledge. Vancouver, British Columbia, Canada; 2012, 1–3.

10. Bienkowski M, Feng M, Means B. *Enhancing teaching and learning through educational data mining and learning analytics: an issue brief*. Washington, D.C.: Office of Educational Technology, U.S. Department of Education; 2012, 1–57.
11. Romero C, Ventura S. *Data Mining in E-learning*. Southampton, UK: Wit-Press; 2006.
12. Castro F, Vellido A, Nebot A, Mugica F. Applying data mining techniques to e-learning problems. In: *Evolution of Teaching and Learning Paradigms in Intelligent Environment*. Studies in Computational Intelligence. Vol. 62. Berlin, Germany: Springer-Verlag; 2007, 183–221.
13. Baker RSJd. Data mining for education. In McGaw B, Peterson P, Baker E, eds. *International Encyclopedia of Education*. 3rd ed. Vol. 7. Oxford, UK: Elsevier; 2010, 112–118.
14. Romero C, Ventura S, Salcines E. Data mining in course management systems: Moodle case study and tutorial. *Comput Edu* 2008, 51:368–384.
15. Zaïane O. Web usage mining for a better web-based learning environment. In: *Proceedings of Conference on Advanced Technology for Education*. Madison, WI; 2001, 60–64.
16. Baker RS, Corbett AT, Koedinger KR, Wagner AZ. Off-task behavior in the cognitive tutor classroom: when students game the system. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vienna, VA; 2004, 383–390.
17. Zaïane O. Building a recommender agent for e-learning systems. In: *Proceedings of the International Conference on Computers in Education*. Auckland, New Zealand; 2002, 55–59.
18. Baker RS, Corbett AT, Koedinger KR. Detecting student misuse of intelligent tutoring systems. In: *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*. Maceiò, Alagoas, Brazil; 2004, 531–540.
19. McCalla G. The ecological approach to the design of e-learning environments: purpose-based capture and use of information about learners. *J Interact Media Edu* 2004, 7:1–23.
20. Sison R, Shimura M. Student modeling and machine learning. *Int J Artif Intell Edu* 1998, 9:128–158.
21. Zaïane O. Towards evaluating learners' behaviour in a web-based distance learning environment. In: *Proceedings of the IEEE International Conference on Advanced Learning Technologies*. 2001, 357–360.
22. Tang T, Mcalla G. Smart recommendation for an evolving e-learning system: architecture and experiment. *Int J E-Learn* 2005, 105–129.
23. Ma Y, Liu B, Wong C, Yu P, Lee S. Targeting the right students using data mining. In: *KDD '00: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2000, 457–464.
24. Marquez-Vera C, Cano A, Romero C, Ventura S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl Intell*. In Press.
25. Merceron A, Yacef K. Mining student data captured from a web-based tutoring tool: initial exploration and results. *J Interact Learn Res* 2004, 15:319–346.
26. Romero C, Zafra A, Luna JM, Ventura S. Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data. *Expert Syst J Knowl Eng*. In Press.
27. Calders To, Pechenizkiy M. Introduction to the special section on educational data mining. *ACM SIGKDD Explor* 2011, 13:3–6.
28. Garcia E, Romero C, Ventura S, Castro C. Collaborative data mining tool for education. In: *International Conference on Educational Data Mining*. Cordoba, Spain; 2009, 299–306.
29. Baker RSJd, Yacef K. The state of educational data mining in 2009: a review and future visions. *J Edu Data Min* 2009, 3–17.
30. Romero C, Espejo P, Zafra A, Romero J, Ventura S. Web usage mining for predicting marks of students that use Moodle courses. *Comput Appl Eng Edu J*. In Press.
31. Baker RSJd, Gowda SM, Corbett AT. Automatically detecting a student's preparation for future learning: help use is key. In: *Fourth International Conference on Educational Data Mining*. Eindhoven, The Netherlands; 2011, 179–188.
32. Vellido A, Castro F, Nebot A. *Clustering Educational Data. Handbook of Educational Data Mining*. Boca Raton, FL: Chapman and Hall/CRC Press; 2011, 75–92.
33. 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*. Beijiing, China; 2004, 248–253.
34. Merceron A, Yacef K. Measuring correlation of strong symmetric association rules in educational data. In Romero C, Ventura S, Pechenizkiy M, Baker RSJd, eds. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press; 2010, 245–256.
35. Rabbany R, Takaffoli M, Zaïane O. Analyzing participation of students in online courses using social network analysis techniques. In: *International Conference on Educational Data Mining*. Eindhoven, The Netherlands; 2011, 21–30.
36. Trčka N, Pechenizkiy M, van der Aalst W. Process mining from educational data. *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press; 2011, 123–142.

37. Tane J, Schmitz C, Stumme G. Semantic resource management for the web: an e-learning application. In: *International Conference of the WWW*. New York; 2004, 1–10.
38. Mazza R, Milani C. GISMO: a graphical interactive student monitoring tool for course management systems. In: *International Conference on Technology Enhanced Learning*. Milan, Italy; 2004, 1–8.
39. Corbett A, Anderson J. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Model User-Adapted Interact* 1995, 4:253–278.
40. Desmarais MC. Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explor* 2011, 13:30–36.
41. Romero C, Ventura S. Preface to the special issue on data mining for personalised educational systems. *User Model User-Adapted Interact* 2011, 21:1–3.
42. Tang T, Daniel BK, Romero C. Preface to the special issue on recommender systems for and in social and online learning environments. *Expert Syst J Knowl Eng*. In Press.
43. Kotsiantis S, Patriarcheas K, Xenos MN. A combinational incremental ensemble of classifiers as a technique for predicting student's performance in distance education. *Knowl-Based Syst* 2010, 23:529–535.
44. Frias-Martinez E, Chen S, Liu X. Survey of data mining approaches to user modeling for adaptive hypermedia. *IEEE Trans Syst Man Cybern C,* 2006, 36, 6, 734–749.
45. Pavlik P, Cen H, Koedinger K. Learning factors transfer analysis: using learning curve analysis to automatically generate domain models. *Int Conf Edu Data Min* 2009, 121–130.
46. Ayers E, Nugent R, Dean N. A comparison of student skill knowledge estimates. In: *International Conference On Educational Data Mining*. Cordoba, Spain; 2009, 1–10.
47. Hsia t, Shie A, Chen L. Course planning of extension education to meet market demand by using data mining techniques—an example of Chinkuo Technology University in Taiwan. *Expert Syst Appl J* 2008, 34:596–602.
48. Wauters K, Desmet P, Noortgate W. Acquiring item difficulty estimates: a collaborative effort of data and judgment. In: *International Conference on Educational Data Mining*. Eindhoven, The Netherlands; 2011, 121–128.
49. Mikut R, Reischl M. Data mining tools. *WIREs: Data Min Knowl Discov* 2011, 1,5, 431–443.
50. Romero C, Ventura S, De Bra P. Knowledge discovery with genetic programming for providing feedback to courseware author. *User Model User-Adapted Interact* 2004, 14:425–464.
51. Merceron A, Yacef K. Educational data mining: a case study. In: *International Conference on Artificial Intelligence in Education*. Amsterdam; 2005, 1–8.
52. Becker K, Vanzin M, Ruiz D. Ontology-based filtering mechanisms for web usage patterns retrieval. In: *Sixth International Conference on E-Commerce and Web Technologies*. Copenhagen, Denmark; 2005, 267–277.
53. Avouris N, Komis V, Fiotakis G, Margaritis M, Voyiatzaki E. Why logging of fingertip actions is not enough for analysis of learning activities. In: *Workshop on Usage Analysis in Learning Systems, AIED Conference*. Amsterdam; 2005, 1–8.
54. 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*. Amsterdam; 2005, 15–22.
55. Bellaachia A, Vommina E. MINEL: a framework for mining e-learning logs. In: *Fifth IASTED International Conference on Web-based Education*. Mexico; 2006, 259–263.
56. Jovanovic J, Gasevic D, Brooks C, Devedzic V, Hatala M. LOCO-Analyist: a tool for raising teacher's awareness in online learning environments. In: *European Conference on Technology-Enhanced Learning*. Crete,Greece; 2007, 112–126.
57. Hershkovitz A, Nachmias R. Developing a log-based motivation measuring tool. In: *First International Conference on Educational Data Mining*. Montreal, Canada; 2008, 226–233.
58. Selmoune N, Alimazighi Z. A decisional tool for quality improvement in higher education. In: *International Conference on Information and Communication Technologies*. Damascus, Syria; 2008, 1–6.
59. Juan A, Daradoumis T, Faulin J, Xhafa F. SAMOS: a model for monitoring students' and groups' activities in collaborative e-learning. *Int J Learn Technol* 2009, 4:53–72.
60. Gaudioso E, Montero M, Talavera L, Hernandez-del-Olmo F. Supporting teachers in collaborative student modeling: a framework and an implementation. *Expert Syst Appl* 2009, 36:2260–2265.
61. Romero C, Ventura S, Zafra A, De Bra P. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Comput Edu* 2009, 53, 828–840.
62. Johnson M, Barnes T. EDM visualization tool: watching students learn. In: *Third International Conference on Educational Data Mining*. Pittsburgh, PA; 2010, 297–298.
63. Pedraza-Perez R, Romero C, Ventura S. A Java desktop tool for mining Moodle data. In: *International Conference on Educational Data Mining*. Eindhoven, The Netherlands; 2011, 319–320.

64. Bakharia A, Dawson S. SNAPP: a bird's-eye view of temporal participant interaction. In: *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*. Vancouver, British Columbia, Canada; 2011, 168–173.
65. Graf S, Ives C, Rahman N, Ferri A. AAT: a tool for accessing and analysing students' behaviour data in learning systems. In: *First International Conference on Learning Analytics and Knowledge*. Banff, Alberta, Canada; 2011, 174–179.
66. Zafra A, Romero C, Ventura S. DRAL: a tool for discovering relevant e-activities for learners. *Knowl Inf Syst*. In Press.
67. García-Saiz D, Zorrilla ME. A service oriented architecture to provide data mining services for non-expert data miners. *Decis Support Syst*. In Press.

