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.

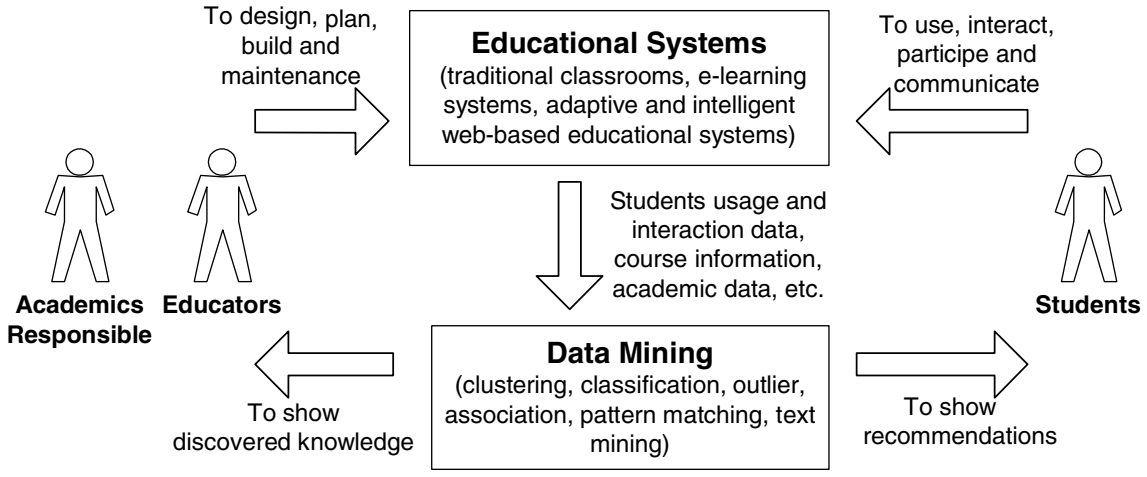

Fig. 1. The cycle of applying data mining in educational systems.