Computers & Education 72 (2014) 23–36

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA

Juan A. Lara [a,*], David Lizcano [a], María A. Martínez [a], Juan Pazos [b], Teresa Riera [c]

[a] Open University of Madrid, UDIMA, Facultad de Enseñanzas Técnicas, Ctra. De la Coruña, Km 38.500, Vía de Servicio, 15, 28400 Collado Villalba, Madrid, Spain
[b] Technical University of Madrid, School of Computer Science, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
[c] Universidad de las Islas Baleares, Departamento de Matemáticas e Informática, Edificio Anselm Turmeda, Crta.Valldemossa Km 7.5, 07122 Palma de Mallorca, Spain

ARTICLE INFO

*Article history:*
Received 18 July 2013
Received in revised form
18 October 2013
Accepted 20 October 2013

*Keywords:*
Educational data mining
Reference model
Educational performance
E-learning
Moodle

ABSTRACT

In today's open and dynamic learning environment, a significant percentage of students have a preference for flexible learning systems whereby they can reconcile their academic pursuits with their job responsibilities and family obligations.
Non face-to-face educational models, like e-learning (*electronic learning*), evolved in order to offer such flexibility. E-learning systems have major strengths but also pose major challenges to the educational community.
One such challenge is the large spatial and temporal gap between the teacher and student, which is an obstacle to student follow-up by teachers. The information generated by virtual learning systems sometimes overwhelms instructors who are unable to process the data without the support of special-purpose techniques and tools that are useful for analysing large dataflows.


In this paper, we propose the use of knowledge discovery in databases (KDD) to extract knowledge that teachers are likely to find useful by analysing data generated by the interaction of students with e-learning environments, like Moodle for example. To do this, our proposal is able to build historical reference models of students that dropped out of and students that completed the course. These generated models can be used to classify a specific student within the dropout or non-dropout group.

Our proposal has been evaluated on real academic data for students enrolled in several courses, generating very satisfactory results with respect to the representativeness of the resulting reference models.

## 1. Introduction

The European Higher Education Area (EHEA) is a higher education initiative implemented as of the 1999 Bologna declaration, whose goal is to set up a common educational space for promoting European development through mobility and increased employment opportunities for European citizens. This necessitated a profound deliberation on the teaching methodologies used to date in order to achieve a common paradigm aiming to offer flexibility to students in response to today's social demands in an open and dynamic environment (Galán & Rodríguez, 2012).

Non face-to-face, primarily web-based educational models, including e-learning (*electronic learning*), are starting to emerge in response to this demand for flexibility. E-learning is a form of distance learning that is completely virtualized through an electronic channel (medium), like the Internet, for example, using learning process support tools.

* Corresponding author. Facultad de Enseñanzas Técnicas, Open University of Madrid, UDIMA, Ctra. De la Coruña, km 38.500, Vía de Servicio, 15, 28400 Collado Villalba, Madrid, Spain. Tel.: +34 630524530.
*E-mail address:* juanalfonso.lara@udima.es (J.A. Lara).

0360-1315/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compedu.2013.10.009

| Guardado en:7 de noviembre de 2012, 11:37 | | | | |
|---|---|---|---|---|
| | | | | |
| Curso | Fecha | Dirección IP | Acción | Información |
| 2012-13_5 | 2012 noviembre 6 23:42 | 93.156.141.60 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 23:13 | 83.52.62.115 | forum view discussion | He subido el .java borrador del ejercicio |
| 2012-13_5 | 2012 noviembre 6 23:12 | 83.52.62.115 | forum view discussion | Dudas sobre implementación |
| 2012-13_5 | 2012 noviembre 6 23:11 | 83.52.62.115 | forum view forum | Foro de tutorias unidad 2 |
| 2012-13_5 | 2012 noviembre 6 23:10 | 83.52.62.115 | forum view discussion | Integracion de Datos (ERP) |
| 2012-13_5 | 2012 noviembre 6 23:10 | 83.52.62.115 | forum view forum | Actividad 4. AA: Foro: novedades y tendencias en integración de datos |
| 2012-13_5 | 2012 noviembre 6 23:10 | 83.52.62.115 | resource view | Guía docente de la asignatura |
| 2012-13_5 | 2012 noviembre 6 23:09 | 83.52.62.115 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 23:09 | 83.52.62.115 | forum view forum | Foro de tutorias Unidad 3 |
| 2012-13_5 | 2012 noviembre 6 23:09 | 83.52.62.115 | choice view | Sesión 3 Elluminate |
| 2012-13_5 | 2012 noviembre 6 23:09 | 83.52.62.115 | choice choose | Sesión 3 Elluminate |
| 2012-13_5 | 2012 noviembre 6 23:09 | 83.52.62.115 | choice view | Sesión 3 Elluminate |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | quiz view | Actividad 3: Control I: UD 1-2 |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | assignment view | Actividad 2. AEC: Caso práctico I |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 21:04 | 88.1.73.95 | forum view forum | Foro de tutorias unidad 2 |
| 2012-13_5 | 2012 noviembre 6 21:03 | 88.1.73.95 | forum view discussion | He subido el .java borrador del ejercicio |
| 2012-13_5 | 2012 noviembre 6 21:03 | 88.1.73.95 | forum view forum | Foro de tutorias unidad 2 |
| 2012-13_5 | 2012 noviembre 6 21:03 | 88.1.73.95 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 20:37 | 186.6.172.172 | forum view discussion | Integracion de Datos (ERP) |
| 2012-13_5 | 2012 noviembre 6 20:37 | 186.6.172.172 | forum view forum | Actividad 4. AA: Foro: novedades y tendencias en integración de datos |
| 2012-13_5 | 2012 noviembre 6 20:37 | 186.6.172.172 | course view | Integración de Bases de Datos - 5301201-1S-B1 |
| 2012-13_5 | 2012 noviembre 6 19:26 | 190.167.211.10 | forum view forum | Actividad 4. AA: Foro: novedades y tendencias en integración de datos |
| ... | | | | |

**Fig. 1.** Extract of an e-learning support tool activity log.

E-learning systems have major strengths. However, they also pose significant challenges to the educational community. We think that one of the major open e-learning problems within the EHEA is the large spatial and temporal gap between teacher and student. This stands in the way of student follow-up by the teacher, which is much more targeted and in step within face-to-face environments. Student supervision and tutoring is essential for detecting student behaviours day to day that can lead to course dropout or poor student preparation for final examinations. To do this, teachers have need of up-to-date, reliable, summarized and easily interpretable information about each student.

It is true that e-learning support tools provide a huge amount of information, basically in the shape of data and news, on student use of and participation in the virtual classroom. Rather than being an aid, this information often overwhelms teachers who are unable to process the hundreds of records generated by student actions in their classrooms. For example, Fig. 1[1] shows an extract of a table containing the student operations log for one day of a course (the student identification field has been omitted on privacy grounds). This table illustrates only a small fraction of student activity for just one course. A full day's table may contain hundreds (if not thousands) of records, which complicates student follow-up.

In face of the huge amount of information that is hard for the teacher to interpret and analyse, it is essential to use special-purpose tools to analyse huge amounts of data and gain as much knowledge as possible. One of the most important approaches for this purpose is knowledge discovery in databases (KDD), which aims to extract useful, implicit and previously unknown knowledge from large volumes of data and/or news. This process ranges from data comprehension and preparation to interpretation and exploitation of the data extraction results. Data mining is a stage within the KDD process where a set of techniques and tools are used to examine and extract useful hidden information from the data (Fayyad, Piatetsky-Shapiro & Smyth, 1996).

We use data mining techniques in this research in order to improve student follow-up by the teacher. The proposal described here, called System for Educational Data Mining (SEDM), is based on the study of two groups of students for each course: students that drop out of the course before taking the final examination and students that do not drop out, pass the continuous assessment and sit the final examination. We analysed these two groups to discover the representative patterns of each group. These patterns are very useful in teaching for classifying a particular student as a member of either of the two above groups and for taking the proper measures improve student performance.

Data mining techniques have been applied to data from e-learning environments over recent years (Baker & Yacef, 2009; Castro, Vellido, Nebot, & Mugica, 2007; Romero & Ventura, 2007; Romero & Ventura, 2010). These techniques address a range of problems, but, as far as we know, none provides a comprehensive framework identifying temporal patterns of behaviour indicative of a significant student underperformance which can then result in a failure to achieve the course learning outcomes.

The remainder of the document is organized as follows. Section 2 describes the baseline e-learning platform used in this research. Section 3 presents the main data mining proposals applied to the field of distance education. Section 4 explains our proposal. Section 5 reports the results of this research. Finally, Section 6 includes the preliminary conclusions and future lines of research.

## 2. Moodle

The baseline system used in this research is Moodle. Moodle is a free virtual learning environment (VLE). As of 2013, Moodle has over 77,000 registered sites in over 225 countries. It provides support to over 66 million students all over the world, tutored by over 1.2 million

[1] The table is Spanish because it is a real extract from an operations log of a course that we teach. Translation of the main elements in Fig. 1.•*Guardado en*: Saved on•*Curso*: Course•*Fecha*: Date•*Dirección IP*: IP Address•*Acción*: Action•*Información*: Information.