UNIVERSIDAD DE CÓRDOBA

Programa de doctorado:

Computación avanzada, energía y plasmas

TÍTULO:

MEJORA EN EL DESCUBRIMIENTO DE MODELOS DE MINERÍA DE PROCESOS EN EDUCACIÓN MEDIANTE AGRUPACIÓN DE DATOS DE INTERACCIÓN CON LA PLATAFORMA MOODLE

Tesis presentada por:

Alejandro Bogarín Vega

Directores:

Dr. D. Cristóbal Romero Morales

Dra. Dª. Rebeca Cerezo Menéndez

Córdoba Julio de 2018

TITULO: *MEJORA EN EL DESCUBRIMIENTO DE MODELOS DE MINERÍA DE PROCESOS EN EDUCACIÓN MEDIANTE AGRUPACIÓN DE DATOS DE INTERACCIÓN CON LA PLATAFORMA MOODLE*

AUTOR: *Alejandro Bogarín Vega*

© Edita: UCOPress. 2018
Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

https://www.uco.es/ucopress/index.php/es/
ucopress@uco.es

UNIVERSITY OF CÓRDOBA

Doctoral Programme:

Advanced computing, energy and plasmas

TITLE:

IMPROVING THE DISCOVERY OF EDUCATIONAL PROCESS MINING MODELS BY GROUPING INTERACTION DATA WITH MOODLE PLATFORM

A Thesis presented by:

Alejandro Bogarín Vega

Advisors:

Dr. D. Cristóbal Romero Morales

Dra. Dª. Rebeca Cerezo Menéndez

Córdoba July, 2018

TÍTULO DE LA TESIS: MEJORA EN EL DESCUBRIMIENTO DE MODELOS DE MINERÍA DE PROCESOS EN EDUCACIÓN MEDIANTE AGRUPACIÓN DE DATOS DE INTERACCIÓN CON LA PLATAFORMA MOODLE.

DOCTORANDO: Alejandro Bogarín Vega

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

El doctorando (Alejandro Bogarín Vega) ha progresado enormemente como investigador desde que en el año 2014 realizara su trabajo de investigación tutelada con los mismos directores y temática, que dio pie a la realización de la actual tesis.

Durante estos 4 años el doctorando ha realizado todas las actividades obligatorias y opcionales (63 en total), trabajado duro seguido siempre las pautas de trabajo que le hemos marcado los directores y el plan de investigación que se estableció.

Como fruto del buen trabajo realizado, de esta tesis se han derivado las siguientes publicaciones:

- 2 Artículo publicado en revista indexada en el JCR (Q2).
- 1 Capítulo de libro indexado en el BCI (Q1).
- 2 Artículo en congreso internacional (Core B).
- 1 Artículo publicado en revista nacional.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 18 de Junio de 2018

Firma del/de los director/es

Fdo.:___Cristóbal Romero Morales___ Fdo.: __Rebeca Cerezo Menéndez_____

La tesis titulada “Mejora en el descubrimiento de modelos de minería de procesos en educación mediante agrupación de datos de interacción con la plataforma Moodle”, que presenta D. Alejandro Bogarín Vega para optar al grado de Doctor, ha sido realizada dentro del programa de doctorado computación avanzada, energía y plasmas, en la línea de investigación aprendizaje automático, modelado de sistemas y minería de datos, del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, bajo la dirección de los doctores Cristóbal Romero Morales y Rebeca Cerezo Menéndez cumpliendo, en su opinión, los requisitos exigidos a este tipo de trabajos.

Córdoba, Julio de 2018

El Doctorando

Fdo: Alejandro Bogarín Vega

El Director

Fdo: Dr. Cristóbal Romero Morales

La Directora

Fdo: Dra. Rebeca Cerezo Menéndez

Esta tesis ha sido parcialmente subvencionada con los proyectos TIN2017-83445-P y EDU2014-57571-P del Ministerio Español de Ciencia, Innovación y Universidades. También se han recibido fondos de la Unión Europea y el Principado de Asturias, a través de su Plan de Ciencia, Tecnología e Innovación (GRUPIN14-053).

AGRADECIMIENTOS

La consecución de esta tesis no ha sido resultado de una sola persona. De una u otra manera, han contribuido un conjunto de personas a la realización de la misma.

Agradezco especialmente a mi directores Dr. Cristóbal Romero y Dra. Rebeca Cerezo todo el apoyo incondicional brindado durante este tiempo. Rebeca, muchas gracias por tus valiosos comentarios en los trabajos realizados con los que he conseguido ser más perfeccionista y por todo el tiempo dedicado. Cristóbal, gracias por confiar en aquel muchacho desconocido que te solicitó un día que le dirigieras su tesis y, el que además de considerarte un gran director, te considera un amigo. Espero y deseo que, aunque nos hayan quitado la tostada de pisto, sigamos desayunando juntos.

A mis queridos padres, que desde los comienzo de mis estudios me han apoyado incondicionalmente y animado a seguir superándome. Gracias por transmitirme los valores de esfuerzo, trabajo y superación.

A mi amada esposa por su amor, consejos, aliento, compresión y por hacerme consciente de que lo que importa en la vida no es lo que te sucede, sino cómo reaccionas a lo que te sucede. Victoria, gracias por tu apoyo, y perdón por el tiempo robado para realizar esta tesis, tiempo que nunca volverá.

Muchas gracias a todos.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

LISTA DE ACRÓNIMOS

AHS	Adaptive Hypermedia System
AM	Alpha Miner
ASR	Association Rule Mining
BPMN	Business Process Model and Notation
CBA	Computer-Based Assessment
CPN	Colored Petri Net
CSCL	Computer-Supported Collaborative Learning
CW	Collaborative Writing
DM	Data Mining
EDM	Educational Data Mining
EM	Esperanza-Maximización
EP	Episode Mining
EPC	Event Process Chain
EPM	Educational Process Mining
ETM	Evolutionary Tree Miner
FM	Fuzzy Miner
GEDM	Graph-Educational Data Mining
GM	Graph Mining
GPL	GNU General Public License
HLPN	High-level Petri Net
HM	Heuristic Miner
IF	Impact Factor
IM	Inductive Miner
ITS	Intelligent Tutoring System
JEDM	Journal of Educational Data Mining

LAK	Learning Analytics and Knowledge
LAS	Lag Sequential Analysis
LM	Learnflow Mining
LMS	Learning Management System
LTL	Linear Temporal Logic
MOOC	Massive Open Online Course
MXML	Mining eXtensible Markup Language
PDA	Personal Digital Assistant
PM	Process Mining
SNA	Social Network Analysis
SOLAR	Society for Learning Analytics Research
SPM	Sequence Pattern Mining
SRL	Self-regulated learning
UML	Unified Modeling Language
VLE	Virtual Learning Environment
WM	Workflow Mining
XES	eXtensible Event Stream

RESUMEN

El desarrollo de la integración entre tecnología y sistemas de aprendizaje, nos permiten capturar todas las acciones que realizan los estudiantes cuando interactúan con los entornos de aprendizaje virtuales. Estas plataformas virtuales de enseñanza almacenan todas las actividades en ficheros o bases de datos que, procesados correctamente, pueden ofrecer información muy útil para la toma de decisiones y responder cuestiones del profesorado en aras de mejorar la calidad del proceso de enseñanza-aprendizaje. Con el objetivo de entender los patrones o rutas seguidas por los estudiantes durante el proceso de aprendizaje, las técnicas de minería de datos en educación están siendo utilizadas de manera exponencial sobre los registros de eventos de estas plataformas. En esta tesis utilizamos técnicas de minería de procesos en educación en sistemas de gestión de aprendizaje, en concreto Moodle, una disciplina emergente y fuertemente relacionada con la minería de datos en educación.

Actualmente, Moodle no proporciona herramientas de visualización específicas de los datos generados por los estudiantes que permitan a los educadores entender esta gran cantidad de información, y tomar consciencia de lo que está pasando durante el proceso de aprendizaje. Por tanto, el objetivo general de esta tesis ha sido descubrir modelos de procesos sobre la interacción de los estudiantes, a partir de los registros de eventos generados por los estudiantes en la plataforma Moodle y que sean generales, visuales, fiables y fáciles de interpretar. Para lograr esta meta, en primer lugar, realizamos un estudio de búsqueda bibliografía sobre minería de procesos en educación. Una vez realizado un estado del arte, propusimos una codificación de alto nivel de los eventos de bajo nivel que proporciona la plataforma Moodle acerca de la interacción de los estudiantes. Además, agrupamos y dividimos los datos de los estudiantes en base a diferentes criterios (nota final y temas). Finalmente, comparamos los diferentes algoritmos de minería de procesos utilizados en educación en base a medidas de calidad.

Los conjuntos de datos utilizados proceden de estudiantes de grado de una universidad del norte de España, a los cuales se les han aplicado varios algoritmos de descubrimiento de minería de procesos junto con diferentes metodologías basadas en técnicas de agrupamiento. Los algoritmos de descubrimiento utilizados son el Alpha Miner, Heuristic Miner, Evolutionary Tree Miner e Indutive Miner. Las metodologías de agrupamiento usadas se hacen de forma manual (por nota y por temas) y automática (variables relacionadas con la interacción de los estudiantes en Moodle). Asimismo, realizamos experimentos agrupando por temas para poder analizar más exhaustivamente el comportamiento de los estudiantes y utilizamos una codificación de alto nivel con cinco

etiquetas con el objetivo de conseguir modelos más comprensibles de acuerdo con los principios teóricos de aprendizaje autoregulado. La herramienta de investigación utilizada para nuestras investigaciones ha sido ProM.

Al realizar el estado del arte sobre la minería de procesos en educación conseguimos conocer cuáles eran los algoritmos y herramientas más utilizadas y con mejores resultados. Se observó que los algoritmos Alpha Miner, Heuristic Miner y Fuzzy Miner eran los algoritmos más utilizados para descubrir modelos de aprendizaje, concretamente Heurisitc Miner era el que mejores resultados mostraba. Posteriormente, descubrimos que con el nuevo algoritmo Inductive Miner se podían obtener mejores resultados que con estos algoritmos tradicionales, incluido Heurisitc Miner. Asimismo, en nuestras investigaciones propusimos, con éxito, diferentes tipos de agrupamientos (manual y automático) para mejorar los modelos de minería de procesos en educación y, al mismo tiempo, optimizar el rendimiento (métricas) y la comprensibilidad (tamaño) de los modelos. Además, se realizaron las pruebas por temas, y conseguimos analizar en mayor profundidad el comportamiento de los estudiantes. Con esta forma de dividir nuestros conjuntos de datos hemos obtenido modelos más específicos. Por otro lado, utilizamos una codificación de alto nivel con cinco etiquetas y obtuvimos un nivel de abstracción mayor y modelos más comprensibles y sencillos desde el punto de vista de los supuestos de la teoría de aprendizaje autoregulado. Finalmente, la utilización de diferentes métricas de evaluación de los modelos obtenidos nos sirvió para contrastar de manera empírica tres importantes conclusiones: en primer lugar, con el algoritmo Inductive Miner obtenemos los mejores resultados en la medida del ajuste. En segundo lugar, los resultados obtenidos en el balanceo de las métricas de calidad (overall) son mejores en el Inductive Miner que en otros algoritmos tradicionales de minería de procesos en educación. Por último, los resultados obtenidos en las métricas, analizadas en conjunto o individualmente, son aún mejores en los conjuntos de datos que estaban agrupados.

ABSTRACT

The development of the integration between technology and learning systems allows us to trace all the actions that students perform when they interact with virtual learning environments. These virtual teaching platforms store all the activities in files or databases that, correctly processed, can provide very useful information for decision making and answer questions from teachers in order to improve the quality of the teaching-learning process. In order to understand the patterns or routes followed by students during the learning process, educational data mining techniques are being implemented exponentially on the event logs of these platforms. In this thesis we use educational process mining techniques on learning management systems, particularly Moodle, an emerging discipline strongly related to educational data mining.

Currently, Moodle does not provide specific visualization tools for the data generated by students that allow educators to understand this large amount of information, and become aware of what is happening during the learning process. Therefore, the general objective of this thesis has been to discover models of processes about students' interactions; going from the records of events generated by students in the Moodle platform to general, visual, reliable, and easily readable models. In order to achieve this goal, firstly, we carried out a bibliography search study on educational process mining. Once a state of the art was carried out, we proposed a high level coding of the low level events that the Moodle platform provides about student interaction. In addition, we grouped and divided student data based on different criteria (final marks and units of knowledge). Finally, we compare the different process mining algorithms used in education based on their quality measures.

Our datasets come from graduate students from a university in the north of Spain, where several process mining discovery algorithms have been applied along with different methodologies based on grouping techniques. The discovery algorithms used are Alpha Miner, Heuristic Miner, Evolutionary Tree Miner, and Inductive Miner. The grouping methodologies used are done manually (by marks and by units of knowledge) and automatically (variables related to the interaction of students in Moodle). Likewise, we carry out experiments grouping by units to be able to analyze more exhaustively the behavior of the students and we used high level coding with five action labels in order to produce more easily understandable models in accordance with assumptions of self-regulated learning. The research tool used for our research has been ProM.

After developing the state of the art about educational process mining we concluded which algorithms and tools were most used and with the best results. It was observed that the algorithms Alpha Miner, Heuristic Miner, and Fuzzy Miner were the algorithms most used to discover learning models, and specifically Heuristic Miner was the one that showed the best results. Later, we discovered that with the new Inductive Miner algorithm, better

results could be obtained than with these traditional algorithms, including Heuristic Miner. Furthermore, we successfully proposed different types of groupings (manual and automatic) to improve educational process mining models and, at the same time, optimize the performance (metrics) and the comprehensibility (size) of the models. In addition, the tests were conducted by units of knowledge, and we managed to analyze in deeper the behavior of the students. With this way of dividing our data sets we have obtained more specific models. On the other hand, we used a high level coding with five labels, and we obtained a higher level of abstraction and more understandable and simple models from the point of view of the self-regulated learning theory. Finally, the use of different evaluation metrics of the models obtained let us to empirically contrast three important conclusions: Firstly, with the Inductive Miner algorithm we obtain the best results in the adjustment measure; secondly, the results obtained in the balancing of the quality metrics (overall) are better in the Inductive Miner than in other traditional educational process mining algorithms; finally, the results obtained in the metrics, analyzed together or individually, are even better in the data set that were grouped.

Parte I. Tesis Doctoral

1

INTRODUCCIÓN

Hoy en día, gracias al desarrollo de la integración entre tecnología y entornos de aprendizaje, los sistemas de información nos permiten capturar todos los eventos que realizan los estudiantes en estos entornos con diferentes niveles de granularidad. Estos eventos pueden ser de bajo nivel, como las pulsaciones de las teclas y los clics de ratón, o de alto nivel, como las rutas de aprendizaje seguidas por los estudiantes (Trcka et al., 2011). El análisis de estos eventos mediante técnicas de minería de datos (Data Mining, DM) en entornos virtuales de aprendizaje (Virtual Learning Environments, VLEs), puede ofrecer información muy útil para la toma de decisiones por parte de alumnos, profesores e instituciones en aras de mejorar la calidad del sistema educativo (Romera & Ventura, 2007).

La minería de datos se define como el descubrimiento de conocimiento para encontrar información no trivial, previamente desconocida y potencialmente útil de grandes repositorios de datos (Frawley et al., 1990). Es un área multidisciplinar donde convergen diferentes paradigmas de la computación como árboles de decisión, inducción de reglas, redes neuronales artificiales y aprendizaje basado en instancias, así como diversos métodos como clasificación, agrupamiento (Dutt et al., 2015; Vellido et al., 2011), y estimación, entre otros.

La aplicación de técnicas de DM a datos recogidos en entornos educativos se denomina minería de datos en educación (Educational Data Mining, EDM) y permite descubrir nuevo conocimiento útil para resolver problemas educativos (Romero & Ventura,

2010). Este nuevo conocimiento, puede ser útil tanto para los profesores como para los estudiantes. A los estudiantes se les puede recomendar actividades y recursos que favorezcan su aprendizaje y, los profesores, pueden conocer el comportamiento que tienen los estudiantes en la plataforma y profundizar en el proceso de aprendizaje que llevan a cabo. De esta manera, un profesor podría adaptar sus cursos al modo en que trabajan sus alumnos y tomar medidas ante los problemas que se puedan detectar.

Debido al interés creciente de esta nueva disciplina, en el año 2008, se celebró en Montreal (Canadá) la primera conferencia específica sobre este temática (International Conference on Eduational Data Mining), se formó el grupo internacional denominado, *The International Working Group on Educational Data Mining*[1] y la revista JEDM (Journal of Educational Data Mining). Algunos años después, en el 2011, apareció otra conferencia (Learning Analytics and Knowledge, LAK), sociedad (Society for Learning Analytics Research, SOLAR) y revista (Journal of Learning Analytics) estrechamente relacionadas con EDM. Aunque EDM y LAK tienen objetivos comunes (Siemens & Baker, 2012), sus diferencias radican en donde cada una de ellas hace más hincapié en sus investigaciones, de forma que EDM da más importancia a los algoritmos de DM utilizados, mientras que LAK da más importancia a los datos y la aplicación final de los resultados.

Tanto EDM como LAK se han aplicado en multitud de tareas educativas, concretamente, en EDM una de las desarrolladas con más éxito ha sido el descubrimiento de patrones, secuencias y rutas de aprendizaje realizadas por los estudiantes dentro de los entornos de educativos (Romero & Ventura, 2017). Sin embargo, debido a que las técnicas clásicas de EDM se centran en descubrir patrones específicos, no proporcionan una representación visual del proceso general que sería de gran ayuda para interpretar estos resultados por parte de los diferentes agentes educativos (Weijters et al., 2006). Para resolver este problema, en los últimos años se está proponiendo el uso de una de las técnicas más prometedoras de EDM, la Minería de Procesos en Educación (Educational Process Mining, EPM).

La Minería de Procesos en Educación (EPM) es una nueva sub-disciplina de EDM que aplica minería de procesos estrictamente a datos educativos (Romero et al., 2016). Tanto EDM como EPM aplican algoritmos específicos a los datos para descubrir patrones y relaciones ocultas, pero a diferencia de EDM, las técnicas de EPM están centradas en el proceso y en los datos del evento (van der Aalst et al., 2004). Además, las técnicas clásicas de EDM son de poca utilidad en el descubrimiento de flujos de control, y no se centran en el proceso de una manera global. Para permitir este tipo de análisis general, en el que el

[1] http://www.educationaldatamining.org

proceso y no el resultado desempeña el papel central, se ha propuesto un nuevo método de investigación de DM, denominado Minería de Procesos (Process Mining, PM).

Estas técnicas de PM son capaces de extraer conocimiento de los registros de eventos disponibles en los sistemas de información actuales, y nos facilitan nuevos medios para descubrir, monitorizar y mejorar los procesos en una gran variedad de dominios de aplicación (van der Aalst, 2011). Hay dos razones principales para el creciente interés en PM. Por un lado, se registran más y más eventos, proporcionando información detallada acerca de la historia de los procesos. Por otro lado, hay una necesidad de mejorar y apoyar los procesos educativos en ambientes competitivos y que cambian rápidamente. PM se puede entender como un puente entre DM y el modelado y análisis de procesos (van der Aalst, 2016). Concretamente PM tiene como principales objetivos (Trcka & Pechenizkiy, 2009):

- Construir modelos completos y compactos de procesos educativos que sean capaces de reproducir todo el comportamiento observado.
- Comprobar si el comportamiento modelado coincide con el comportamiento observado.
- Proyectar información extraída de los registros en el modelo para hacer explícito el conocimiento tácito y facilitar una mejor comprensión del proceso.

A destacar algunas de las aplicaciones de PM en educación (Bogarín et al., 2018a), como:

- Ayudar a una mejor comprensión de los procesos educativos.
- Descubrir las rutas de aprendizaje realizadas por los estudiantes.
- Generar recomendaciones y consejos a los estudiantes.
- Proporcionar una retroalimentación a los estudiantes, profesores y/o investigadores.
- Detectar problemas de aprendizaje temprano.
- Ayudar a los estudiantes con alguna dificultad de aprendizaje.
- Mejorar la gestión de los objetivos de aprendizaje.

Finalmente, de entre todas las anteriores, se puede destacar una de gran actualidad e importancia: comprender como los estudiantes interactúan y aprenden dentro de entornos de aprendizaje muy demandantes cognitiva y metacognitivamente, como los hipermedia y, descubrir que rutas siguen (Azevedo et al., 2012). Esta línea es el punto de partida de la actual tesis doctoral.

1.1 Objetivos

El **objetivo general** de esta tesis es descubrir modelos de procesos sobre la interacción (rutas de aprendizaje seguidas) de los estudiantes, a partir de los registros de eventos (ficheros logs) generados por los estudiantes en la plataforma Moodle y que sean generales, visuales, fiables y fáciles de interpretar.

Los siguientes **objetivos específicos** se han marcado para lograr esta meta:

- **O_1:** Realizar un estudio de búsqueda bibliografía sobre minería de procesos en educación.

- **O_2:** Proponer una codificación de alto nivel de los eventos de bajo nivel que proporciona la plataforma Moodle acerca de la interacción de los estudiantes.

- **O_3:** Agrupar y dividir los datos de los estudiantes en base a diferentes criterios, e.g. por nota final obtenida en el curso, o por temas en los que se divide la asignatura.

- **O_4:** Comparar los diferentes algoritmos de minería de procesos utilizados en educación en base a medidas de calidad.

1.2 Hipótesis

Nuestras hipótesis de partida para los objetivos planteados han sido:

- **H_1:** Si llevamos a cabo un estado del arte sobre esta nueva disciplina de minería de procesos educativos, lograremos conocer cuáles son los algoritmos y herramientas más utilizadas y con mejores resultados.

- **H_2:** Si codificamos los ficheros de datos proporcionados por Moodle utilizando en lugar de los eventos de bajo nivel, una nomenclatura de más alto nivel semántico que nos proporcione un nivel de abstracción superior de las diferentes acciones realizadas por los alumnos, será más sencillo interpretar los modelos obtenidos.

- **H_3:** Si agrupamos los datos utilizando diferentes criterios y los dividimos en varios ficheros de datos (en lugar de utilizar todo el conjunto de datos), podremos obtener modelos más específicos que sean además más certeros y comprensibles, y

evitaremos modelos demasiado amplios y complejos para ser interpretados por un profesor.

- **H_4:** Si comparamos los diferentes algoritmos de descubrimiento de modelos de procesos utilizando varias medidas de calidad, podremos determinar que algoritmo o algoritmos descubren los mejores modelos que describen el comportamiento o rutas de los estudiantes en un curso de Moodle.

1.3 Propuesta

En esta tesis se propone la aplicación de técnicas de EPM sobre los datos de la interacción de los estudiantes con un sistema de gestión de aprendizaje (Learning Management System, LMS), específicamente Moodle, con el objetivo de descubrir modelos que proporcionen información útil a profesores e investigadores sobre el comportamiento de los estudiantes dentro de dicha plataforma.

En este sentido, la obtención de modelos que proporcionan una representación visual compresible para los profesores ha sido una de las principales contribuciones de esta tesis. Los resultados generados pueden ser útiles para el seguimiento del aprendizaje de los estudiantes y para proporcionar una retroalimentación a profesores y alumnos, con la que se pueda tomar consciencia de lo que está pasando durante el proceso de aprendizaje.

Los conjuntos de datos que se van a utilizar proceden de una institución educativa de nivel superior (Universidad de Oviedo). Se han recogido datos de Moodle durante varios cursos académicos para una asignatura del grado en Psicología. Para el procesamiento de estos datos se han utilizado herramientas específicas como Microsoft Access y Excel. Una vez pre-procesados los datos, se han aplicado varios algoritmos de descubrimiento de minería de procesos, junto con varias metodologías basadas en técnicas de agrupamiento. Para ello, la herramienta software de EPM utilizada para la realización de todos los experimentos ha sido ProM (van der Aalst, 2011), desarrollada por la Universidad Técnica de Eindhoven[2] y distribuida con licencia GPL (GNU General Public License).

[2] https://www.tue.nl/

1.4 Estructura

La figura 1.1 muestra la particular estructura que sigue esta tesis, en la que hay dos bloques fundamentales. En el primero se resume la tesis doctoral en los apartados de introducción, marco teórico, objetivos, metodología, resultados y conclusiones. En la segunda parte se aportan los dos artículos publicados en revista científicas internacionales con índice de impacto (Impact Factor, IF), el capítulo del libro publicado en la editorial WILEY, los congresos internacionales y revistas nacionales.

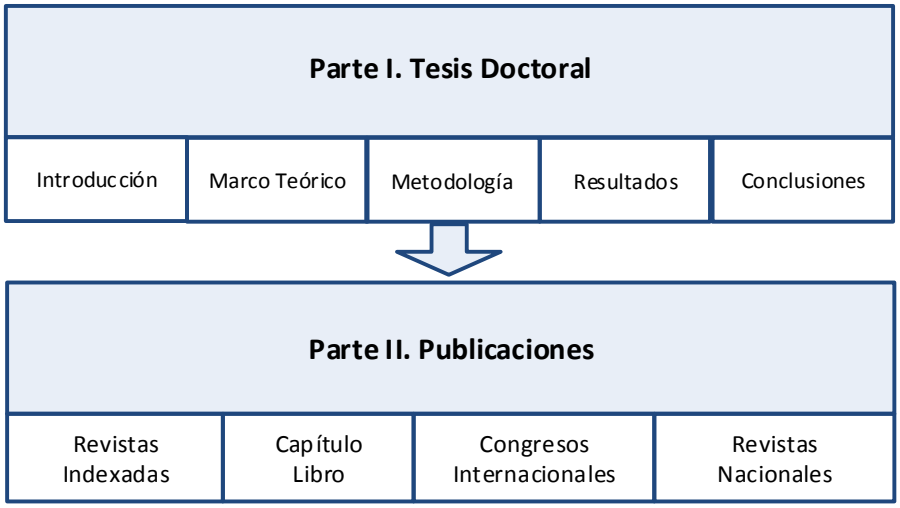


Figura 1.1: Esquema general de la tesis.

.

2

MARCO TEÓRICO

En este capítulo se realiza un estudio bibliográfico exhaustivo del EPM. Se describen los dominios de aplicación más relevantes de la disciplina, se detallan los componentes principales del marco de EPM y se abordan los principales obstáculos encontrados cuando realizamos el tratamiento de datos de los registros de eventos obtenidos de entornos educativos. Asimismo, se detallan cómo son los datos utilizados, herramientas, técnicas y modelos más usados en EPM. Finalmente, se presenta una visión general de los principales trabajos de investigación realizados hasta el momento en esta disciplina, agrupados por dominios de aplicación.

2.1 Áreas relacionadas

PM es una tecnología relativamente nueva que surge dentro de la comunidad empresarial (van der Aalst et al., 2004). Se centra en el desarrollo de técnicas dirigidas a extraer conocimiento relacionado con los procesos de los registros de eventos. Utiliza los ficheros que se registran en los sistemas de información para descubrir, supervisar y mejorar procesos en diferentes dominios, así como para verificar la conformidad de procesos, detectar cuellos de botella y predecir problemas. La mayoría de los trabajos de PM se han centrado en el descubrimiento de flujos de trabajo a través de representaciones con redes de Petri (Trcka & Pechenizkiy, 2009). Estos métodos toman la información de los registros de eventos como entrada produciendo modelos de procesos que describen la

información de los registros de una manera global (Reimann et al., 2014). PM también se conoce como minería de flujo de trabajo (Workflow Mining, WM) o minería de flujo de aprendizaje (Learnflow Mining, LM), que en conexión con WM, ha sido utilizado por algunos autores como Bergenthum et al. (2012) o Perez-Rodriguez et al. (2009), mientras que muchos otros (Cairns et al., 2015a; Romero & Ventura, 2013; van der Aalst et al., 2013) prefieren el término EPM en relación con la minería de procesos en educación. Asimismo, hay otras metodologías de investigación relacionadas que se han utilizado para descubrir el comportamiento de los estudiantes (ver tabla 2.1). A continuación, abordaremos brevemente tres de las que están más estrechamente relacionados con PM: Minería de Intención, Minería de Patrones Secuenciales (Sequence Pattern Mining, SPM) y Minería de Grafos (Graph Mining, GM).

Minería de intención

La minería de intención es un campo de investigación vinculado con PM que pretende determinar la intención que subyace a la conducta del usuario en base a los registros de su interacción con un sistema informático, como por ejemplo, en búsquedas realizadas en motores de búsqueda. Un conjunto de acciones se corresponde con el logro de una intención; al igual que PM, la minería de intención utiliza registros de eventos como entrada y produce modelos de procesos intencionales, entendiendo intención como la determinación a actuar en un cierto camino sentido (Khodabandelou et al., 2013).

Es importante señalar que no hemos encontrado ninguna investigación sobre la aplicación de IM al campo de la educación, pero el potencial de esta técnica puede ser fácilmente intuido debido a que es particularmente adecuada para estudiar las llamadas learning intentions.

Minería de patrones secuenciales

SPM (Agrawal & Srikant, 1995) es una técnica muy utilizada en el entorno de la minería de datos para descubrir sub-secuencias frecuentes entre varios o muchos usuarios. El análisis secuencial de patrones tiene como objetivo encontrar si existe algún orden específico dentro de los casos (Nesbit et al., 2007). SPM está relacionado con la Minería de Episodios (Episode Mining, EP); de hecho, ambas técnicas pueden ser vistas como variantes de la Asociación de Minería de Reglas (Association Rule Mining, ASR). Sin embargo, los métodos SPM encuentran los patrones de eventos más frecuentes a lo largo de un conjunto de secuencias de eventos, mientras que EP descubre los patrones de eventos más frecuentemente utilizados dentro de una secuencia dada. Existen otras técnicas

relacionadas con SPM, como Lag Sequential Analysis (LAS), análisis de t-pattern y modelos de Markov. Todas estas técnicas son más adecuadas para secuencias recurrentes relativamente cortas y análisis de transiciones de eventos (Reimann et al., 2009).

Las técnicas de SPM han sido muy aplicadas para analizar los comportamientos de aprendizaje de los estudiantes. Sin embargo, están más indicadas cuando se trata de descubrir patrones de comportamiento más simples que un proceso. Por lo tanto, SPM no es apropiado para descubrir comportamientos de aprendizaje que abordan el proceso de aprendizaje de manera global (Bannert et al., 2014).

Minería de grafos

GM es otra técnica popular de minería de patrones. El objetivo de GM es encontrar todos los sub-gráficos frecuentes en un gráfico mayor o una base de datos de gráficos. GM y DM están estrechamente relacionados. El primero es más orientado a la geometría y el segundo más orientado a la lógica y la relación (Washio & Motoda, 2003). También es importante diferenciar entre GM y Análisis de Redes Sociales (Social Network Analysis, SNA); SNA puede ser considerado como una aplicación de GM.

La minería de datos educativos basada en gráficos (Graph Educational Data Mining, GEDM) es también una nueva área de investigación relacionada. Tanto GEDM como EPM utilizan gráficos para representar la información. Sin embargo, mientras que la tarea de GM es extraer patrones a través de gráficos que describen los datos subyacentes (sub-gráficas de interés) y podrían ser utilizados más, por ejemplo, para la clasificación o agrupación, PM se centra en el proceso de manera global y por lo tanto sus gráficos descubren el proceso general de aprendizaje. Respecto a esto, cabe destacar que los gráficos son extremadamente importantes en la comunidad EDM, ya que muchos tipos de datos pueden representarse como gráficos, incluyendo datos de redes sociales y discusiones online.

Finalmente, en la tabla 2.1 se muestra una comparación de las áreas de investigación EPM previamente descritas.

Tabla 2.1: Principales áreas relacionadas con EPM

	Objetivos	**Algoritmos**	**Modelos**	**Herramientas**
Minería de Procesos	Descubrir los procesos subyacentes en los registros de eventos	Heuristic Miner, Fuzzy Miner, etc.	Petri Nets, Heuristic Net, BMMN, etc.	ProM, Disco, Celonis, etc.
Minería de Intención	Modelar los procesos según el propósito de los actores	Viterbi Algorithm, Baum-Welch Algorithm, etc.	KAOS, I*, Map, etc.	Ninguna herramienta encontrada
Minería de Patrones Secuenciales	Encontrar patrones comunes entre los ejemplos de datos donde los valores se entregan en una secuencia	Generalized Sequential Patterns (GSP), Sequential Pattern Mining (SPAM), PrefixSpan, etc.	Secuencias y subsecuencias, reglas	SPFM, Himalaya Data Mining, etc.
Minería de Grafos	Extraer patrones (sub-gráficas) de interés de los gráficos que describen los datos subyacentes	Branch-and-bound, On-line Plan Recognition, Recursive Matrix (R-MAT), etc.	Probabilistic graphs, signed graphs, colored graphs, Transition graphs, etc.	Graphviz, Deep Though, GSLAP, etc.

2.2 Marco y conceptos

En la figura 2.1 se muestra una visión general de la aplicación de PM en el campo educativo. Este marco de EPM es una adaptación del marco genérico de PM (Pechenizkiy et al., 2009) al campo de la educación (Cairns et al., 2015a; Vidal et al., 2016) que no puede entenderse sin la descripción de los principales agentes implicados:

- **Proceso de enseñanza-aprendizaje o universo educativo**. Básicamente, dos actores desempeñan un papel importante en cualquier actividad de formación online: profesores y estudiantes. Los profesores proveen los recursos apropiados para asegurar el éxito de los estudiantes. Los estudiantes son la parte esencial de cualquier actividad de formación online, interactuando con otros participantes (estudiantes o profesores), y con el propio sistema. Finalmente, los cursos,

conferencias, exámenes, etc. simplemente se utilizan como recursos para los participantes.

- **Entorno virtual de aprendizaje**. El entorno donde se desarrolla el proceso de enseñanza-aprendizaje proporciona las estructuras y recursos básicos en los que se producen las acciones de instrucción y las interacciones de los participantes. La mayoría de estos entornos proporcionan a los profesores o investigadores algunas herramientas básicas para analizar el aprendizaje de los estudiantes (evolución de las notas, número de actividades realizadas, participación en el foro, último acceso, etc.), pero no instrumentos específicos que permitan a los educadores evaluar de una manera exhaustiva el proceso general de aprendizaje del estudiante.
- **Registros de eventos**. Los registros son fichero que recopilan los eventos que se producen en los entornos virtuales de aprendizaje y, que normalmente, se almacenan en bases de datos. Contienen una gran cantidad de datos en bruto sobre la interacción de los agentes educativos en el entorno de aprendizaje virtual. Estos registros necesitan ser transformados en un formato de archivo específico para poder ser utilizados por herramientas específicas de PM.
- **Modelos de proceso**. Estos modelos revelan información valiosa sobre cómo los participantes del mundo educativo interactúan con el sistema a partir de los registros de eventos. Se obtienen utilizando diferentes técnicas para descubrir procesos relevantes para el aprendizaje. Se pueden distinguir tres tipos principales de PM (ver tabla 2.2): descubrimiento, conformidad y extensión. Estos tres tipos básicos de PM también se pueden explicar en términos de entrada y salida (ver figura 2.2).

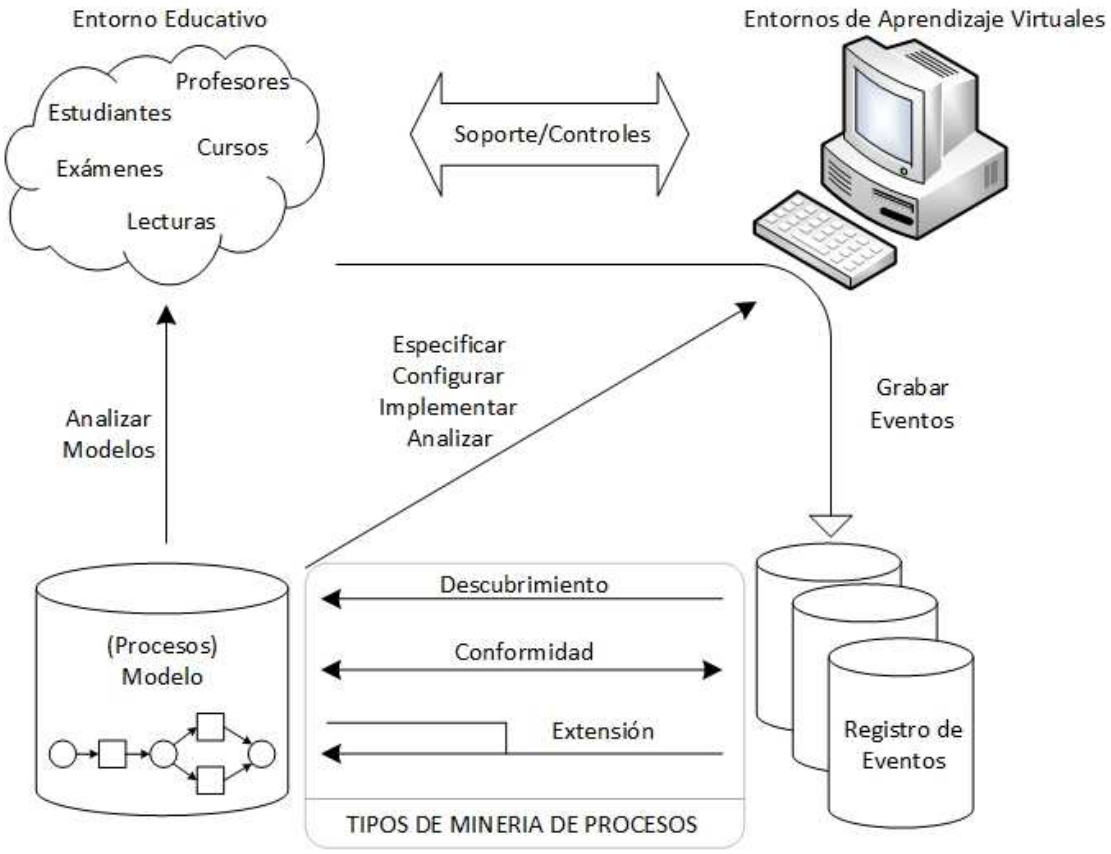


Figura 2.1: Marco EPM: Tipos y componentes.

Tabla 2.2: Tipos de minería de procesos.

Tipo	Descripción	Aplicación en Educación
Descubrimiento de procesos	Construye un modelo de proceso completo capaz de reproducir el comportamiento visto en el archivo de registro.	El profesor puede visualizar el modelo de conducta de los caminos de aprendizaje de los estudiantes, proporcionando conocimiento del proceso en lugar de sólo el resultado del aprendizaje.
Comprobación de conformidad	Encuentra desviaciones entre los comportamientos observados en los registros de eventos y los modelos de procesos generados.	El profesor puede analizar si el modelo obtenido (manual o automático) se corresponde con el modelo de comportamiento de los registros de eventos y, por ejemplo, encontrar valores atípicos.
Extensión o mejora	Tiene como objetivo mejorar o ampliar un modelo de proceso dado, basándose en la información extraída de un registro de eventos específico que está relacionado con el mismo proceso.	El profesor puede detectar cuellos de botella o relaciones entre estudiantes de un curso, ya que diferentes enfoques pueden fusionarse en un único modelo de proceso integrado y extendido.

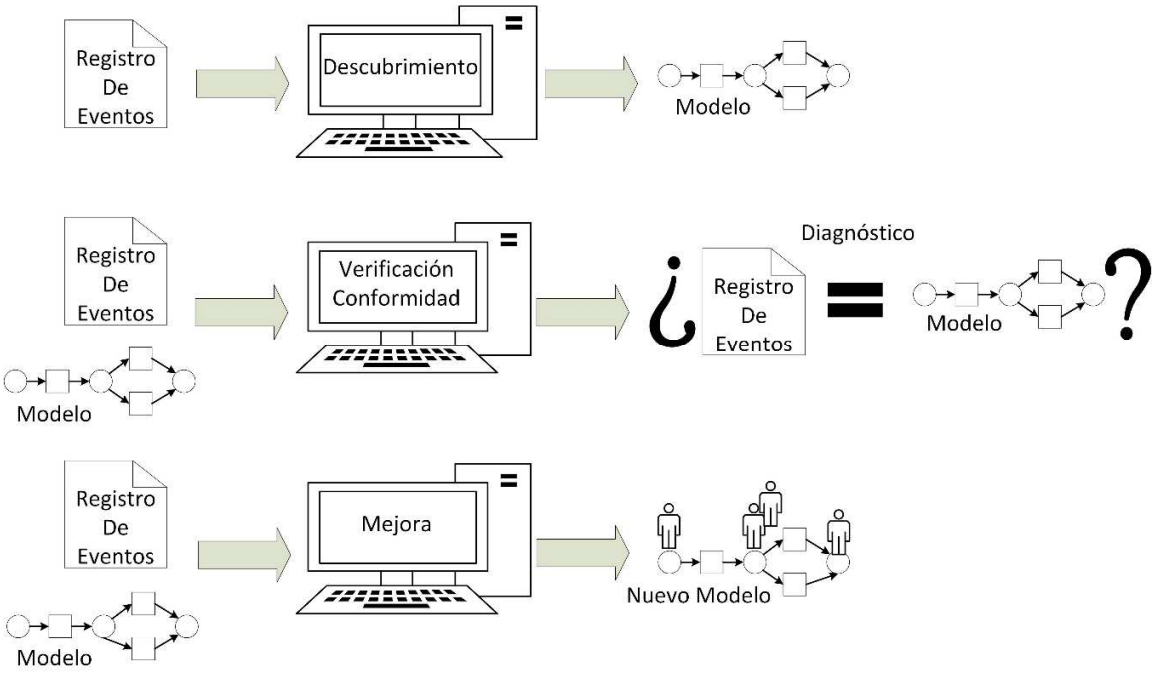


Figura 2.2: Tipos de Minería de Procesos explicados en términos de entrada y salida.

.

Además de los tres tipos principales de PM, PM también proporciona perspectivas distintas (van der Aalst, 2016): de control-flujo, de organización, de caso y de tiempo. La más utilizada en el entorno educativo es la perspectiva de control-flujo que se centra en el ordenamiento de las actividades. El objetivo principal de esta perspectiva es descubrir una descripción ideal de todos los caminos o rutas de aprendizaje imaginables (Schoonenboom et al., 2007) que se pueden generar cuando los estudiantes navegan a través de un entorno virtual de aprendizaje.

2.3 Datos y herramientas

En esta sección se muestra una descripción más detallada de los datos, los diferentes retos encontrados al realizar el tratamiento de los mismos, y las soluciones de software utilizadas para abordar su análisis a través de EPM.

El punto de partida para PM es un registro de eventos (van der Aalst, 2016). Un registro de eventos puede ser una hoja de cálculo de Excel, una tabla de base de datos o un archivo simple que contiene una traza/secuencia de eventos. Cada evento es una fila en el registro de eventos y se refiere a un caso (identificación de caso), una actividad (nombre de actividad) y un punto en el tiempo (marca de tiempo), y en ocasiones puede contener

información adicional. Generalmente, estos ficheros necesitan ser transformados en formatos específicos tales como XES (eXtensible Event Stream) o MXML (Mining eXtensible Markup Language) para poder ser utilizados por una herramienta de PM (Romero et al., 2016). Existen algunas herramientas específicas, como ProMimport, que proporcionan la conversión de diferentes fuentes de datos a estos formatos (van der Aalst, 2016).

Los registros de eventos educativos se pueden recopilar de una amplia gama de entornos virtuales de aprendizaje, tales como los LMSs, Cursos Online Masivos Abiertos (Massive Open Online Courses, MOOCs), Sistemas de Tutoría Inteligentes (Intelligent Tutoring Systems, ITSs), Sistemas Adaptativos de Hipermedia (Adaptive Hypermedia Systems, AHSs), etc. La figura 2.3 muestra un ejemplo de un registro de eventos generado por Moodle (LMS). El sistema Moodle registra en cada clic lo que los diferentes agentes educativos realizan durante la navegación, generando una gran cantidad de información, a priori, sin sentido.

Time	IP Address	Full Name	Action	Information
24/09/2013 12:22	150.214.10.12	Student53	resource view	Tema 1
24/09/2013 12:24	150.214.10.12	Student53	resource view	Tema 2
24/09/2013 12:25	150.214.10.12	Student53	resource view	Tema 3
24/09/2013 12:26	150.214.10.12	Student53	resource view	Tema 4
24/09/2013 12:28	150.214.10.12	Student53	folder view	Clases Prácticas (Diagnóstico)
24/09/2013 12:30	180.45.67.44	Student42	resource view	Tema 1
24/09/2013 12:31	180.45.67.44	Student42	resource view	Tema 2
24/09/2013 12:31	180.45.67.44	Student42	resource view	Tema 3
24/09/2013 13:29	155.123.23.14	Student35	folder view	Clases Prácticas (Diagnóstico)
24/09/2013 13:29	155.123.23.14	Student35	resource view	Tema 1
24/09/2013 13:29	155.123.23.14	Student35	resource view	Tema 2
24/09/2013 13:29	155.123.23.14	Student35	resource view	Tema 3
24/09/2013 13:29	155.123.23.14	Student35	resource view	Tema 4
24/09/2013 14:06	145.124.25.65	Student49	folder view	Clases Prácticas (Diagnóstico)
24/09/2013 14:33	154.132.45.66	Student7	folder view	Clases Prácticas (Diagnóstico)
24/09/2013 14:33	154.132.45.66	Student7	resource view	Tema 1
24/09/2013 14:33	154.132.45.66	Student7	resource view	Tema 2
24/09/2013 14:33	154.132.45.66	Student7	resource view	Tema 3
24/09/2013 14:33	154.132.45.66	Student7	resource view	Tema 4

Marca de Tiempo

Identificador del Caso

Nombre de la Actividad

Información Adicional

Figura 2.3: Ejemplo del registro de eventos de Moodle.

En general, aparecen varios problemas al realizar el tratamiento en los registros de eventos que necesitan ser abordados y tenidos en cuenta para el EPM (Cairns et al., 2015a;

van der Aalst, 2016). En la tabla 2.3 se describen algunos de los problemas más frecuentes y se ilustran con un ejemplo.

Tabla 2.3: Desafíos y problemas al manejar los registros de eventos.

Problema	Descripción	Ejemplo en EPM
Correlación	Los eventos se agrupan por caso en un registro de eventos. Los eventos deben estar relacionados entre sí.	Los estudiantes realizan tipos de acciones similares en un foro.
Ruido	Un registro de eventos puede contener valores atípicos. El comportamiento excepcional no es representativo del comportamiento típico del proceso.	Los estudiantes pueden salir de una sesión abierta.
Imperfección	El registro de eventos contiene muy pocos eventos para poder descubrir algunas de las estructuras de control-flujo subyacentes.	Los sistemas que dan soporte a la formación online fallan, por ejemplo, se cae un servidor.
Distribución	Los datos pueden proceder de más de una fuente de información distinta, de forma que se encuentren distribuida y no centralizada.	La información del estudiante se puede recolectar de diversas fuentes: información administrativa, clases de teoría y de práctica, entornos de aprendizaje online, etc.
Marca de tiempo	Los eventos deben estar ordenados por caso.	Problemas típicos: sólo fechas, zonas horarias diferentes, registro atrasado.
Instantánea	Los casos pueden tener una vida que es anterior o se extiende más allá del período registrado.	Un estudiante inició su actividad antes del inicio del registro de eventos.
Ámbito o Alcance	¿Cuál es el proceso que queremos investigar? ¿Cómo decidir qué tablas incluir?	LMS y MOOC pueden proporcionar diferentes tablas para investigar diferentes procesos.
Granularidad	Los eventos en el registro están a un nivel diferente de granularidad.	La información en educación puede tener diferentes niveles de granularidad: clics de bajo nivel, actividades, cursos, etc.
Contextualización	Los eventos ocurren en un contexto particular que puede explicar ciertos fenómenos. Esto requiere la fusión de los datos de eventos con datos contextuales.	Los profesores descubren modelos en una clase de repetidores.
Tamaño	El número de casos o eventos en los registros de eventos puede ser alto. Estos archivos pueden ser difíciles de manejar debido a su tamaño.	Los entornos virtuales de aprendizaje pueden generar ficheros de importante dimensiones.

Complejidad	Distintas trazas y actividades en los registros de eventos pueden ser de alta complejidad debido a la gran diversidad de comportamientos en los caminos de aprendizaje de los estudiantes.	Los entornos virtuales de aprendizaje pueden generar modelos complejos que son difíciles de entender (espaguetis).
Concept drift	Situación en la que el proceso cambia mientras se analiza.	Los cursos y currículos pueden ser modificados en cualquier momento durante el período de aprendizaje.
Privacidad	La privacidad y autenticación tiene muchas dimensiones éticas.	Los estudiantes necesitan ser conscientes de lo que el sistema está haciendo con sus datos.

Por último, han surgido muchas herramientas para dar soporte a las técnicas de minería de procesos (van der Aalst, 2016): ProM, Disco, Celonis Discovery, Perceptive Process Mining, QPR ProcessAnalyzer, Aris Análisis de Procesos de Negocio, Fujitsu Process Analytics, XMAnalyzer, StereoLOGIC Discovery Analyst, etc., todas ellas son herramientas de PM de uso general y sólo unas pocas han sido usadas para EPM. En la tabla 2.4 se ofrece una comparación entre ellas

Tabla 2.4: Comparación entre las principales herramientas utilizadas en EPM.

	ProM	**Disco**	**SoftLearn**
Compañía (País)	Universidad Técnica de Eindhoven (Holanda)	Fluxicon (Holanda)	Universidad de Santiago de Compostela (España)
Propósito	General	General	Específica (Educación)
Tipo	Gratis	Comercial	Privada
Filtrado	SI	SI	NO
Descubrimiento de procesos	SI	SI	SI
Comprobación de conformidad	SI	NO	NO
Minería de Redes Sociales	SI	NO	NO
Número de Artículos EPM	21	7	1

Sólo tres de estas herramientas de PM han sido referenciadas en el subconjunto de bibliografía relacionada con EPM (ver tabla 2.4). La herramienta ProM, utilizada en esta tesis, es un software genérico de código abierto para implementar PM y, es la más completa

y usada en EPM, seguida por Disco, que también es una herramienta de propósito general pero comercial. Sólo hay un software de PM específico para el dominio educativo, llamado SoftLearn (Barreiros et al., 2014) que proporciona una interfaz gráfica que los profesores pueden utilizar para visualizar rutas de aprendizaje como gráficos de actividad, y así acceder a los datos relevantes generados en las actividades de aprendizaje.

2.4 Técnicas

En esta sección, describimos las técnicas más utilizadas en EPM. Destacamos cuatro grupos principales de técnicas: descubrimiento, verificación de conformidad, análisis de gráfica de puntos y análisis de redes sociales.

Técnicas de descubrimiento

Las técnicas de descubrimiento de procesos construyen un modelo de proceso basado únicamente en un registro de eventos que captura el comportamiento visto en dicho registro; se centran en la perspectiva de control-flujo del proceso. Hay un buen número de algoritmos en PM para descubrir procesos subyacentes en los registros de eventos, pero los más utilizados en los dominios educativos son:

- **Alpha algorithm:** una técnica relativamente intuitiva y sencilla basada en la relación de dependencias entre eventos. Requiere un registro ideal de eventos sin ruido y fue uno de los primeros algoritmos que pudo abordar la concurrencia (Mekhala, 2015).
- **Heuristic Miner algorithm:** utiliza la probabilidad calculando las frecuencias de las relaciones entre las tareas (por ejemplo, dependencia causal, bucles, etc.) y construye tablas de dependencia / frecuencia y gráficas de dependencia / frecuencia (Khodabandelou et al., 2013). El algoritmo Heuristic Miner fue diseñado para hacer uso de una métrica basada en la frecuencia y por lo tanto es menos sensible al ruido y a la imperfección de los registros (Bogarín et al., 2014).
- **Genetic algorithm:** proporciona modelos de procesos basados en matrices causales (dependencias de entrada y salida para cada actividad). Este enfoque aborda problemas como el ruido, datos incompletos, actividades ocultas, concurrencia y actividades duplicadas (Khodabandelou et al., 2013).
- **Fuzzy miner:** es uno de los algoritmos más recientes de descubrimiento de procesos. Es el primer algoritmo que aborda directamente los problemas con un gran número de actividades y un comportamiento altamente no estructurado (Günther & van der Aalst, 2007).

Es necesaria una buena notación para representar los modelos de proceso al usuario final. Todos los algoritmos mencionados anteriormente producen un modelo de proceso que es normalmente independiente de la representación deseada. Existen diferentes tipos de representaciones en PM: redes de Petri, redes de flujo de trabajo, redes difusas, redes heurísticas, redes causales, árbol de procesos, BPMN (Business Process Model and Notation), EPC (Event Driven Process Chain) y diagrama de actividades UML (Unified Modeling Language). Aunque las redes de Petri y BPMN son las más utilizadas en PM (Khodabandelou et al., 2013), las más usadas en el ámbito de la educación son (ver tabla 2.5):

- **Redes de Petri:** gráficos con dos tipos de nodos enlazados por arcos dirigidos. El primer tipo de nodo se conoce como lugar y está representado por una elipse. Los lugares pueden almacenar un conjunto múltiple de valores, denominados tokens. El segundo tipo de nodo, las transiciones, se representan con rectángulos e identifican elementos activos de la red (Vidal et al., 2012).
- **Red de Petri de Alto Nivel (High-level Petri Net, HLPN):** Redes de Petri clásicas pero ampliadas con color, tiempo y jerarquía. Las redes de Petri de color (Colored Petri Nets, CPN) fueron la primera materialización concreta de HLPN y fueron un lenguaje gráfico para analizar las propiedades de los sistemas concurrentes (Khodabandelou et al., 2013).
- **Fuzzy net:** simplifican el modelo completo manteniendo las aristas o eventos altamente significativos, agregando y agrupando las aristas y nodos menos significativos pero altamente correlacionados, y abstrayendo las aristas y nodos menos significativos y mal correlacionados, eliminándolos del modelo simplificado (Günther & van der Aalst, 2007).
- **Redes heurísticas**: Un gráfico de ciclo dirigido que representa los comportamientos más frecuentes de los estudiantes en el conjunto de datos utilizado. En las redes heurísticas las casillas cuadradas representan las acciones y los arcos / enlaces representan dependencias / relaciones entre acciones (Bogarín et al., 2014).

Además, es posible transformar automáticamente un modelo de una representación a otra cuando se utilizan herramientas potentes de PM. En la figura 2.4 mostramos dos representaciones diferentes obtenidas del mismo registro de eventos. Una red de Petri que muestra la causalidad y el paralelismo de los eventos y, una red heurística que muestra la frecuencia de los eventos y cómo de fuerte es la dependencia entre estos eventos.

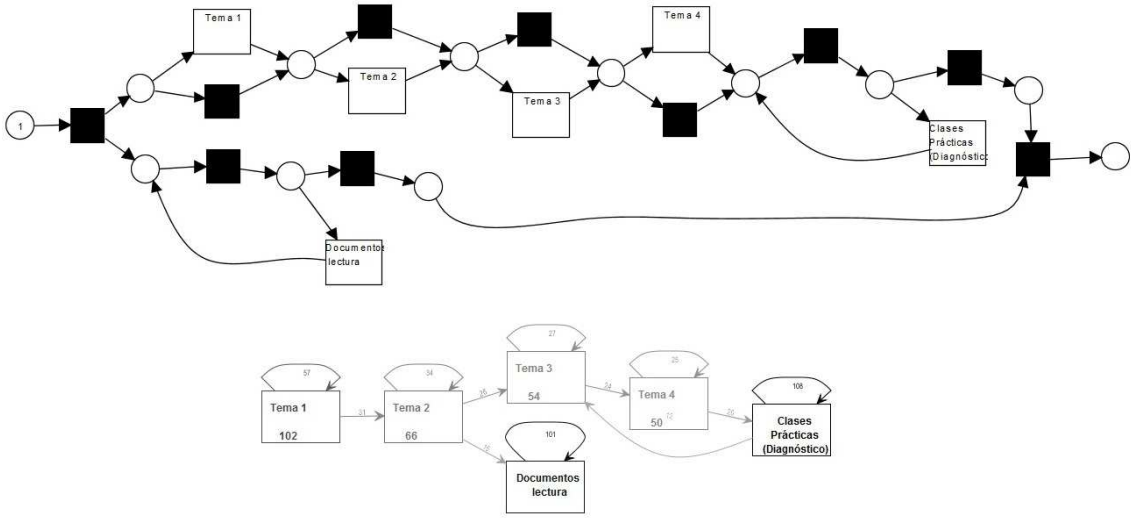


Figura 2.4: Ejemplos de Red de Petri y Red Heurística generados con los mismos datos de registro.

En este sentido, se puede afirmar que el modelo de representación más utilizado en las investigaciones de EPM es la red Fuzzy, seguida por la red de Petri y la red heurística, siendo HLPN el menos usado (ver tabla 2.5 para obtener información detallada).

Tabla 2.5: Modelos de representación utilizados en los trabajos de EPM.

Trabajo/Paper	PETRI NETS	HLPN	FUZZY	HEURISTIC
Weijters et al., 2006	X			X
Günther & van der Aalst, 2007			X	
Pechenizkiy et al., 2009	X		X	X
Reimann et al., 2009	X			X
Trcka & & Pechenizkiy, 2009		X		
Southavilay et al., 2010				X
Trcka et al., 2011	X		X	
Poncin et al., 2011a			X	

Schoor & Bannert, 2012			X	
Anuwatvisit et al., 2012	X			
Ayutaya et al., 2012	X			X
Bergenthum et al., 2012	X	X		
van der Aalst et al., 2013			X	
Reimann et al., 2014			X	
Bannert et al., 2014	X		X	
Cairns et al., 2014b				X
Cairns et al., 2014a			X	
Bogarin et al., 2014				X
Cairns et al., 2015b	X			X
Cairns et al., 2015a			X	X
Mukala et al., 2015b			X	
Ariouat et al., 2016				X
Doleck et al., 2016			X	
Okoye et al., 2016			X	
Sedrakyan et al., 2016			X	
Vahdat et al., 2015			X	
Vidal et al., 2016	X			

Técnicas de comprobación de conformidad

El objetivo de la comprobación de la conformidad es encontrar coincidencias y discrepancias entre el comportamiento modelado y el comportamiento observado. En la literatura de EPM, dos técnicas destacan en la verificación de conformidad:

- Verificador de Lógica Temporal Lineal (Linear Temporal Logic, LTL), que comprueba si los registros de eventos satisfacen alguna fórmula de lógica temporal lineal (LTL) (Van Dongen et al., 2005). El verificador LTL no compara un modelo con el registro, sino con un conjunto de requisitos descritos por LTL.
- El verificador de conformidad (Conformance Checker), que requiere un modelo además de un registro de eventos. Reproduce un registro de eventos en un modelo de red de Petri mientras reúne información de diagnóstico a la que se puede acceder posteriormente (Rozinat & van der Aalst, 2005).

Técnica de análisis de puntos

Un gráfico de puntos muestra la propagación de los eventos a lo largo del tiempo trazando un punto para cada evento de un registro de eventos y, proporcionando así una idea del proceso subyacente, su rendimiento y cualquier patrón de interés. Representa el archivo de registro visualmente, mostrando una perspectiva temporal del proceso de una manera general. El gráfico tiene dos dimensiones ortogonales: el tiempo y los tipos de componentes. El tiempo se mide a lo largo del eje horizontal del gráfico, los tipos de componentes se muestran a lo largo del eje vertical (Cairns et al., 2015b). La figura 2.5 muestra un ejemplo de gráfico de puntos del trabajo diario realizado por los estudiantes en Moodle. Cada fila es una tarea diferente de Moodle en el curso y, el tamaño de los puntos representa cuántos estudiantes han hecho esta tarea en un momento determinado.

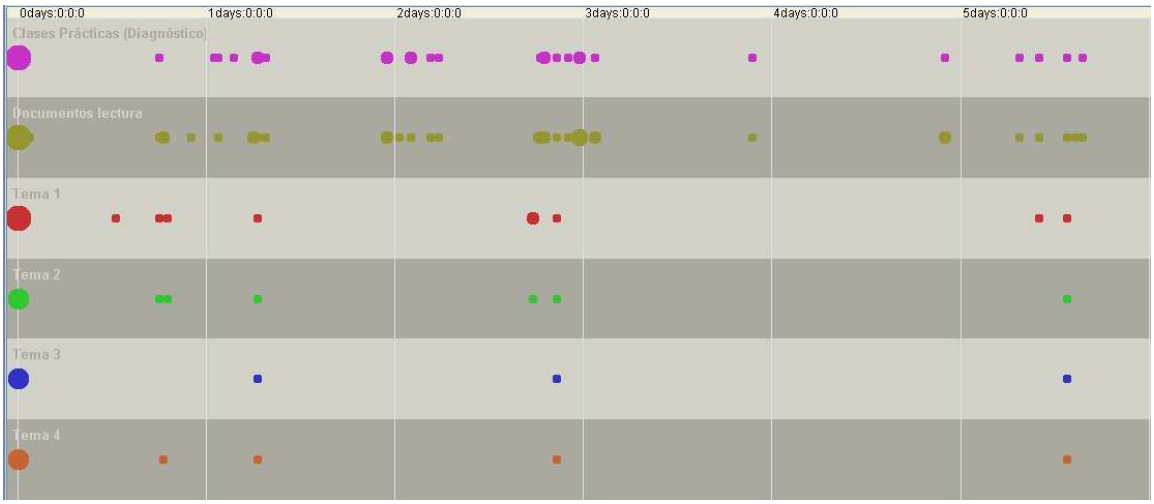


Figura 2.5: Ejemplo de un gráfico de puntos del trabajo diario realizado por los estudiantes en Moodle.

Técnica de análisis de redes sociales

El Análisis de Redes Sociales se refiere a la recopilación de métodos, técnicas y herramientas de sociometría orientadas al análisis de redes sociales. SNA pretende extraer las redes sociales de los registros de eventos basándose en las interacciones observadas entre los participantes, dependiendo de cómo las instancias del proceso se orientan entre estos participantes (Cairns et al., 2014a). Una red social consiste en nodos que representan entidades de una organización y arcos que representan relaciones. La figura 2.6 muestra un ejemplo de redes sociales que representan cómo y cuánto interactúan los estudiantes en un foro de Moodle. Los nodos más grandes representan a estudiantes más activos y los arcos representan el momento en que interactúan.

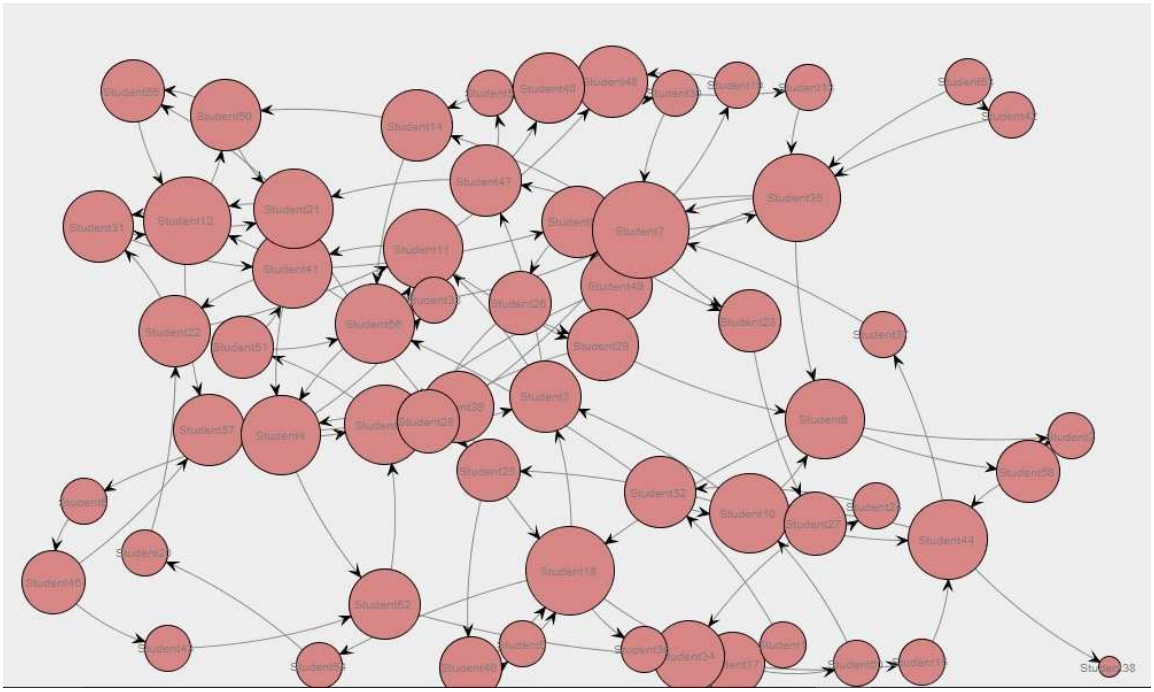


Figura 2.6: Ejemplo de una red social que representa cómo y cuánto interactúan los estudiantes en un foro de Moodle.

Por último, la tabla 2.6 muestra un resumen de las técnicas de descubrimiento, representación y comprobación de conformidad más utilizadas en investigaciones de EPM.

Tabla 2.6: Técnicas utilizadas en investigaciones de EPM.

Trabajo/Paper	**Algoritmo de Descubrimiento**	**Técnicas de Conformidad**	**Cuadro de Puntos**	**SNA**
Weijters et al., 2006	Heuristic Miner			
Pechenizkiy et al., 2009	Heuristic Miner Fuzzy Miner	Conformance Checker	X	
Reimann et al., 2009	Heuristic Miner			
Trcka & Pechenizkiy, 2009		Conformance Checker		
Southavilay et al., 2010	Heuristic Miner		X	
Trcka et al., 2011	Fuzzy Miner	LTL- Conformance Checker	X	
Poncin et al., 2011a	Fuzzy Miner		X	

Ayutaya et al., 2012	Heuristic Miner			
Anuwatvisit et al., 2012		Conformance checker		
Schoor & Bannert, 2012	Fuzzy Miner			
van der Aalst et al., 2013	Fuzzy Miner	Conformance Checker	X	
Reimann et al., 2014	Fuzzy Miner			
Barreiros et al., 2014	Genetic Algorithm			
Bannert et al., 2014	Fuzzy Miner	LTL- Conformance Checker		
Cairns et al., 2014b	Heuristic Miner	LTL		
Cairns et al., 2014a	Fuzzy Miner			X
Bogarin et al., 2014	Heuristic Miner			
Cairns et al., 2015b		LTL- Conformance Checker	X	
Cairns et al., 2015a	Fuzzy Miner	LTL- Conformance Checker	X	X
Mukala et al., 2015b	Fuzzy Miner	Conformance Checker	X	
Vahdat et al., 2015	Fuzzy Miner			
Ariouat et al., 2016	Heuristic Miner			
Okoye et al., 2016	Fuzzy Miner			
Sedrakyan et al., 2016	Fuzzy Miner		X	
Vidal et al., 2016	Genetic Algorithm			

De la tabla 2.6 se puede inferir que los algoritmos de descubrimiento más utilizados son Heuristic Miner y Fuzzy Miner. El verificador de conformidad es la técnica de conformidad más comúnmente utilizada y, las gráficas de puntos se utilizan más que el análisis de redes sociales en las investigaciones de EPM.

2.5 Dominios de aplicación

EPM se ha utilizado en una amplia gama de dominios educativos con el fin de abordar diversos problemas; en esta sección, se aborda la literatura más destacada al respecto.

Entornos MOOC, AHS y LMS

MOOC, AHS, LMS y otros entornos similares de aprendizaje online proporcionan oportunidades de aprendizaje gratuitas a una gran comunidad de internautas. Los archivos de registro generados por estos sistemas proporcionan, entre otras cosas, una idea de cómo los participantes siguen el curso, cuando ven, por ejemplo, videos o conferencias, o cuando entregan actividades.

Hay mucha investigación sobre la aplicación de PM en este tipo de entornos de aprendizaje. Trcka et al. (2011) ilustraron la aplicabilidad de PM al extraer conocimiento de los LMSs teniendo en cuenta sólo las trazas de los exámenes de los estudiantes. En Bogarin et al. (2014), los autores utilizaron los datos de los registros de Moodle y propusieron usar clustering para poder obtener modelos de proceso más precisos y específicos del comportamiento de los estudiantes. En un entorno similar, Reiman et al. (2014) propusieron el uso de trazas para estudiar el Aprendizaje Auto-Regulado (Self-Regulated Learning, SRL) en un entorno hipermedia basado en métodos teóricos y de PM. Utilizando estos métodos, Bannert et al. (2014) detectaron diferencias en las frecuencias de eventos de SRL utilizando técnicas de PM y, encontraron que los estudiantes que tenían éxito mostraban más eventos de aprendizaje y uniformes. En otra investigación Mukala et al. (2015a) utilizaron técnicas de PM para rastrear y analizar los hábitos de aprendizaje de los estudiantes basándose en los datos MOOC. Los resultados indicaron que los estudiantes con éxito siguen un patrón secuencialmente estructurado mientras que los estudiantes sin éxito son impredecibles y tienen procesos mal estructurados. En una investigación posterior Mukala et al. (2015b) hicieron uso de la verificación de conformidad para extraer y analizar los patrones de aprendizaje de los estudiantes en un MOOC. Siguiendo una línea similar, Emond & Buffett (2015) aplicaron técnicas de descubrimiento de minería de procesos y técnicas de minería de clasificación de secuencias para modelar y apoyar el SRL en entornos heterogéneos. Por último, Vidal et al. (2016) utilizaron registros de un entorno de aprendizaje virtual para extraer la estructura del flujo de aprendizaje utilizando PM.

Aprendizaje colaborativo asistido por ordenador

El Aprendizaje Colaborativo Asistido por Ordenador (Computer-Supported Collaborative Learning, CSCL) se caracteriza por compartir y construir conocimiento entre los participantes que usan la tecnología como principal medio de comunicación.

PM se ha aplicado en CSCL con el fin de proporcionar una retroalimentación a los estudiantes en sus procesos de toma de decisiones. En Reimann et al. (2009), el objetivo fue utilizar PM para identificar los modelos de los grupos que tomaban decisiones y, que tuvieron lugar en una sala de chat. En un estudio similar, Bergenthum et al. (2012) propusieron un lenguaje de modelado para los flujos de aprendizaje colaborativo que tenía en cuenta específicamente los agentes implicados, los roles y la representación explícita de los grupos. Su investigación se nutre de trabajos previos centrados en el descubrimiento de estructuras para el control de flujo utilizando métodos del área de WM (Bergenthum et al., 2008). Otros autores como Schoor & Banner (2012) han explorado secuencias de procesos de regulación social durante una tarea CSCL y lo han relacionado con el rendimiento del grupo. Este estudio utilizó PM para identificar los patrones del proceso de pares con rendimiento grupal alto y bajo. En una investigación más reciente en este campo, Porouhan & Premchaiswadi (2017) aplicaron varias técnicas de PM como minería de redes sociales y análisis de gráfica de puntos con el objetivo de aumentar el conocimiento del profesor sobre la dinámica colaborativa en cada grupo.

Una aplicación particular de EPM a este dominio es la escritura colaborativa (Collaborative Writing, CW). La CW es ampliamente utilizada en entornos educativos, los estudiantes usan los ordenadores para tomar apuntes durante las clases o escribir redacciones y trabajos. Gracias a la disponibilidad de Internet, los estudiantes también pueden escribir de manera colaborativa compartiendo y editando sus documentos de varias maneras. PM se ha utilizado en Southavilay et al. (2010) para analizar los procesos de escritura de los estudiantes y cómo estos procesos se relacionan con la calidad y características semánticas del producto final. En este estudio se utilizaron documentos recogidos de diferentes grupos de estudiantes universitarios que escribían de manera colaborativa para evaluar las heurísticas propuestas (Boiarsky, 1984) y se ilustra la aplicabilidad de las técnicas de PM para analizar el proceso de escritura.

Formación profesional

Las instituciones han trabajado para que sus cursos de formación profesional sean más cada vez más ágiles para responder a las necesidades cambiantes del mercado de

trabajo y satisfacer los requisitos de tiempo en la adquisición de habilidades profesionales (Cairns et al., 2014a).

PM se ha utilizado en diferentes tipos de formación profesional. Cairns et al. (2014a) mostraron cómo se pueden utilizar PM para monitorizar y mejorar los procesos educativos en este nivel educativo en concreto. El objetivo de su investigación fue desarrollar métodos genéricos que puedan aplicarse a cuestiones de educación general y aplicaciones más específicas en materia de formación profesional o aprendizaje online para la extracción, análisis, mejora y personalización de procesos educativos. En una investigación similar, Cairns et al. (2015b) analizaron los procesos de formación y su cumplimiento con respecto a algunas restricciones establecidas en el currículo y los requisitos previos de los educadores. Su objetivo era intentar mejorar los modelos de los procesos de formación. Para ello, utilizaron tanto indicadores como el tiempo de ejecución, como la detección de cuellos de botella y puntos de decisión. Doleck et al. (2016) aplicaron técnicas de descubrimiento de PM con el objetivo de proporcionar una visión más coherente del razonamiento del diagnóstico clínico en un entorno de aprendizaje médico e informatizado. Vahdat et al. (2016) aprovecharon las técnicas de PM para investigar y comparar los procesos de aprendizaje de estudiantes de formación profesional midiendo la comprensibilidad de los modelos obtenidos usando una métrica de complejidad. Por último, Ariouat et al. (2016) trataron de identificar las mejores rutas de formación utilizando bases de datos de una empresa de consultoría global.

Minería del plan de estudios

Un plan de estudios es parcialmente diseñado por una institución educativa para lograr ciertos objetivos. Los planes de estudio sugieren normalmente que los estudiantes sigan caminos diferentes debido al enfoque libre en la elección de asignaturas (Wang y Zaïane, 2015).

Trcka & Pechenizkiy (2009) propusieron utilizar como guía de ayuda al profesor un conjunto de plantillas que se podían predefinir. De esta manera, se podía enfocar la minería de procesos y hacerla más eficaz y eficiente con el objetivo de poder ayudar a los educadores a analizar y modelar el curriculum académico. En otra investigación relacionada, Wang & Zaïane (2015) descubrieron un modelo de proceso curricular de estudiantes que realizaban diferentes cursos. Compararon las rutas que los estudiantes con éxito y con menos éxito tendían a tomar, resaltando las discrepancias entre ellos. En otro trabajo Schulte et al. (2017) presentaron una investigación sobre minería de procesos en educación y el análisis de los datos de estudiantes universitarios con el objetivo de descubrir patrones estadísticamente importantes y significativos en la elección de su plan de estudios.

Evaluación basada en ordenadores

La evaluación basada en ordenadores (Computer-Based Assessment, CBA) es, en esencia, la práctica de realizar cuestionarios y exámenes a través del ordenador en lugar de usar los formatos tradicionales de lápiz y papel; esta técnica es ampliamente utilizada en muchos entornos de aprendizaje virtual.

En este sentido, PM ha sido utilizado para analizar los datos de evaluaciones procedentes de estudios online con exámenes de elección múltiple, que muestran la utilidad del descubrimiento de procesos, la comprobación de la conformidad y las técnicas de análisis de rendimiento (Pechenizkiy et al., 2009). En un contexto similar, Tóth et al. (2017) describieron cómo extraer información de los registros de eventos y, cómo usar estos datos en evaluaciones de resolución de problemas.

Inscripción de estudiantes

La inscripción de estudiantes se ocupa de todos los requisitos y diferentes fases del proceso de registro académico. Es fundamental comprobar los procesos del sistema de gestión en el ámbito educativo con el fin de producir resultados esperados en estas gestiones en términos de calidad y tiempo (Ayutaya et al., 2012).

En este contexto, Ayutaya et al. (2012) utilizaron el algoritmo Heuristics Miner (HM) para conocer mejor los procesos de registro de estudiantes en una universidad tailandesa. La característica más importante del HM es su robustez contra el ruido y las excepciones. Debido a que HM se basa en la frecuencia de los patrones es posible centrarse en el comportamiento principal del registro de eventos y lo hace especialmente apropiado para los procesos educativos no estructurados. Anuwatvisit et al. (2012) usaron la verificación de conformidad para detectar discrepancias entre los flujos previstos en un modelo de registro de estudiantes y las instancias de proceso reales.

Repositorios de Software

Los desarrolladores y los equipos de desarrollo están involucrados en procesos de desarrollo de software, a menudo, desde diferentes lugares. En estos proyectos se utilizan diferentes tipos de repositorios de software como sistemas de gestión de código fuente, repositorios de documentos, archivos de correo, controladores de errores y sistemas de control de versiones para apoyar la comunicación y la coordinación.

PM también se ha aplicado para minar repositorios de software. Poncin et al. (2011a) identificaron los desafíos que deben ser abordados para permitir esta aplicación. Analizaron cómo se puede tratar y presentar a través de un marco para analizar software de repositorios (Framework for Analyzing Software Repositories, FRASR). Asimismo, Poncin et al. (2011b) ha utilizado PM para describir el proceso de análisis de datos de repositorios de software. La etapa de pre-procesamiento extrae la información desde los diferentes repositorios de software (los cuales tienen estructuras diferentes) y combina esta información en un único registro de eventos. Por otro lado, la etapa del análisis está dirigida a descubrir la estructura del proceso reflejada en el registro y visualizarlo o analizar si es correcto.

Ciclo de investigación estructurado

Un ciclo de investigación estructurado es una estrategia de adaptación del proceso de enseñanza-aprendizaje que combina estructuración explícita y andamiaje, sin renunciar a una experiencia de aprendizaje más libre y personalizada, estando especialmente indicada para aprendices con alta variabilidad de conocimientos previos, habilidades metacognitivas y motivación. Por ejemplo, en educación para adultos online, donde la libertad de navegación, unido a un escaso conocimiento previo del dominio o pobres habilidades de aprendizaje, puede tener efectos negativos en las experiencias de aprendizaje. Howard et al. (2010) mostraron modelos de proceso con Redes de Petri que contribuyeron a la planificación colaborativa y la revisión de los resultados; y en un contexto similar, Jeong et al. (2010) utilizaron un modelo de Markov para estudiar las conductas de aprendizaje de alumnos nobeles en un campo, implementando estrategias de ciclo de investigación estructurado.

Mundos virtuales educativos en 3D

Los Mundos Virtuales Educativos 3D son entornos que fomentan la interacción entre estudiantes y profesores. Estos entornos animan a realizar actividades de aprendizaje que no fueron programadas inicialmente por los profesores, por ejemplo, a través de avatares.

PM también se ha utilizado para descubrir qué está sucediendo en los procesos de aprendizaje de un estudiante dentro de un mundo virtual 3D. Con este objetivo, Fernández-Gallego et al. (2013) presentaron un marco analítico de aprendizaje para mundos virtuales educativos 3D que se centraba en el descubrimiento de flujos de aprendizaje y la verificación de la conformidad a través de técnicas de PM. Hay que destacar que en este dominio específico, se producen una gran cantidad de interacciones entre los estudiantes y

el entorno, produciendo una generación continua de eventos de bajo nivel, muchos de los cuales se pueden catalogar de información ruidosa. En otras palabras, hay un gran número de eventos que no son significativos desde el punto de vista pedagógico y que generarían modelos excesivamente grandes y complejos, por lo que o bien no hay que tenerlos en cuenta o bien agruparlos dentro de actividades de más alto nivel semántico.

Para cerrar este capítulo, en la tabla 2.7 se muestra un resumen de las investigaciones de EPM descritas anteriormente y su objetivo, agrupadas por dominio de aplicación. Por un lado, podemos ver que, actualmente, las investigaciones más activas pertenecen a los dominios de entornos MOOC, AHS y LMS, aprendizaje colaborativo asistido por ordenador y formación profesional. Por otro lado, observamos que los resultados de EPM pueden ser utilizados para comprender mejor los procesos educativos subyacentes, proporcionar retroalimentación a los estudiantes, profesores e investigadores, detectar dificultades de aprendizaje y ayudar a los estudiantes con dificultades de aprendizaje específicas, mejorar la gestión de las metas de aprendizaje, o generar consejos a los estudiantes, entre otras muchas aplicaciones. En lo que respecta a los objetivos, los más frecuentes se centran en comprender mejor los procesos educativos subyacentes, detectar las dificultades de aprendizaje y descubrir los flujos de aprendizaje de los estudiantes (ver tabla 2.7).

Tabla 2.7: Principales estudios publicados, objetivos abordados y dominios de aplicación del EPM.

Aplicación	Trabajo/Paper	Objetivo
Entornos MOOC, AHS y LMS	Mukala et al., 2015b	Detectar dificultades de aprendizaje
	Mukala et al., 2015a	Generar recomendaciones o consejos para los estudiantes.
	Bogarin et al., 2014	Obtener una mejor comprensión del proceso educativo subyacente
	Vidal et al., 2016	Mejorar la gestión de los objetos de aprendizaje
	Bannert et al., 2014	Detectar dificultades de aprendizaje y descubrir patrones secuenciales
	Reimann et al., 2014	Descubrir patrones secuenciales
	Trcka et al., 2011	Descubrir los flujos de aprendizaje
	Emond & Buffett., 2015	

Aprendizaje colaborativo asistido por ordenador	Reimann et al., 2009	Descubrir los flujos de aprendizaje y proporcionar retroalimentación
	Bergenthum et al., 2012	Descubrir los flujos de aprendizaje
	Schoor & Bannert, 2012	Descubrir patrones secuenciales
	Porouhan & Premchaiswadi, 2017	Para obtener una mejor comprensión del proceso educativo subyacente
	Southavilay et al., 2010	Conocer mejor el proceso educativo subyacente y detectar las dificultades de aprendizaje
Formación profesional	Cairns et al., 2014a	Analizar redes sociales
	Cairns et al., 2015b	Descubrir los flujos de aprendizaje
	Doleck et al. 2016	Conocer mejor el proceso educativo subyacente y detectar las dificultades de aprendizaje
	Vahdat et al., 2015	
	Ariouat et al., 2016	Obtener una mejor comprensión del proceso educativo subyacente
Minería del plan de estudios	Trcka & Pechenizkiy, 2009	Obtener una mejor comprensión del proceso educativo subyacente
	Wang & Zaïane, 2015	Obtener una mejor comprensión del proceso educativo subyacente y generar recomendaciones o consejos para los estudiantes.
	Schulte et al., 2017	Generar recomendaciones o consejos para los estudiantes.
Evaluación basada en ordenadores	Pechenizkiy et al., 2009	Proporcionar retroalimentación
	Tóth et al., 2017	Detectar dificultades de aprendizaje
Inscripción de estudiantes	Anuwatvisit et al., 2012	Obtener una mejor comprensión del proceso educativo subyacente

	Ayutaya et al., 2012	Descubrir los flujos de aprendizaje
Repositorios de Software	Poncin et al., 2011b	Obtener una mejor comprensión del proceso de desarrollo de software
	Poncin et al., 2011a	
Ciclo de investigación estructurado	Howard et al., 2010	Detectar dificultades de aprendizaje
	Jeong et al., 2010	
Mundos virtuales educativos en 3D	Fernández-Gallego et al., 2013	Descubrir los flujos de aprendizaje

3

METODOLOGÍA

La metodología que se ha seguido en esta tesis doctoral ha consistido, principalmente, en tres fases o etapas (ver Figura 3.1). En la primera etapa se ha realizado un trabajo de búsqueda y revisión bibliográfica de EPM, en la segunda etapa se han recogido y pre-procesado los datos procedentes del entorno Moodle, y en la última etapa se han ejecutado y comparado diferentes algoritmos de descubrimiento de modelos.

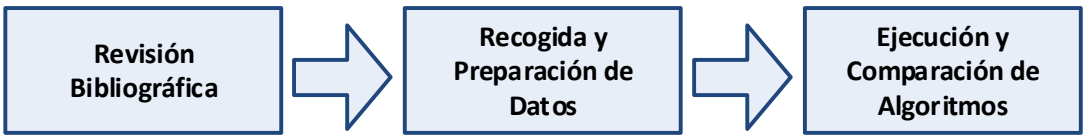


Figura 3.1: Metodología seguida en esta tesis.

3.1 Revisión bibliográfica

Con el objetivo de analizar las técnicas, algoritmos y herramientas más utilizadas, se realizó un estudio bibliográfico de EPM (Bogarín et al., 2018a).

Asimismo, en este estudio concreto, nos propusimos comparar EPM con otras áreas de conocimiento para contrastar las semejanzas y diferencias más significativas. Por otro lado, y con la finalidad de adaptar el marco de trabajo genérico de la minería de procesos a uno específico de educación, desarrollamos y describimos los componentes de este nuevo marco. Además, debido a la dificultad que supone el pre-procesado de datos en entornos educativos, explicamos todos los obstáculos encontrados al manipular los diferentes registros de eventos generados por estos entornos.

Finalmente, categorizamos por ámbito de aplicación todas las investigaciones de EPM realizadas hasta la fecha con el objetivo de ordenar y estructurar de una manera eficaz toda esta información.

3.2 Recogida y pre-procesado de datos

En las investigaciones de Romero et al. (2016) y Bogarín et al. (2018b), los datos utilizados fueron obtenidos de la interacción que realizaron estudiantes universitarios del Grado de Psicología de la Universidad de Oviedo en la plataforma Moodle. La asignatura era de carácter obligatorio en tercero de carrera y constaba de 11 temas diferentes disponibles semanalmente para los estudiantes durante un periodo de 15 días. La interacción de los estudiantes en Moodle genera un registro de eventos que fue procesado de una manera exhaustiva y rigurosa para poder obtener modelos educativos válidos. Los atributos que se utilizaron en nuestras metodologías fueron: la fecha, el nombre del alumno, la acción y un campo información que proporcionaba un conocimiento más detallado de la acción realizada (ver tabla 3.1).

Tabla 3.1: Atributos del registro de eventos de Moodle.

Atributos	**Descripción**
Curso	El nombre del curso
Dirección IP	La IP del dispositivo usado para acceder
Tiempo	La fecha de acceso
Nombre Completo	El nombre del estudiante
Acción	La acción que realiza el estudiante
Información	Más información sobre la acción

Como la mayoría de los registros de eventos del mundo real, los ficheros utilizados en nuestras investigaciones tuvieron que ser filtrado porque contenían ruido (Romero et

al., 2008). Se han tenido que eliminar registros duplicados, registros de profesores, administradores y usuarios de prueba, registros de mensajes de error, etc. Asimismo, se realizó un filtrado de acciones, eliminando los registros que contienen acciones que se consideraban irrelevantes en el rendimiento y calificación final de los estudiantes. Acciones como *ver todos los usuarios, ver todas las etiquetas, ver todas las carpetas*, etc., son eliminadas.

Para nuestros experimentos se utilizaron tres tipos de ficheros. En primer lugar, el registro de eventos obtenido directamente de Moodle y correctamente filtrado (ver figura 2.3). En segundo lugar, un fichero resumen con variables calculadas a partir de la interacción de cada estudiante en Moodle (registro de eventos) y diferentes tablas de bases de datos (ver figura 3.2). Finalmente, usamos un archivo proporcionado por el profesor que contenía un identificador para cada estudiante con el objetivo de mantener su anonimato.

```
@data
0,3.36,11.73,8.73,8.82,4.64,6.18,4.09,77,2.89,5.74,3.2,3.5,2.5,3.63,4.5,5,5,4.13,3.63,4.4,0,cluster1
1,8.45,25.73,21.91,0.18,1.45,2.27,15.91,199.1,6,6.84,4,6,6.25,5.25,5.83,5,1.8,3.75,3.88,7.1,3,cluster3
2,6.45,13.27,10.18,1.64,2.82,3.27,7.55,147.44,4.78,8.47,4.5,7,5.25,6.88,6.67,6.75,1.8,5.75,5.88,9.15,4,cluster3
3,6.36,11.27,3.09,6,5.27,5.55,0,0,0,8.21,3.8,5,3.25,4.5,4.67,6.75,3.6,6.38,5.88,6,5,cluster0
4,1.45,9.55,4.55,0.55,1.82,2.36,2.91,49.29,2.57,6.05,3.2,4.75,4.75,5.13,3.83,4.75,3.2,5.63,3.88,7.25,6,cluster1
5,4.27,10.45,13.36,1.73,1.73,2.91,9.18,165.2,5.4,8,4.3,4.75,6.75,6,5.67,6.25,3.6,6.13,5,8.05,7,cluster3
6,5.91,21.36,7.64,1.91,1.91,2.73,7,68,2.5,9.32,4.6,5.5,6.25,6.25,5.83,5.5,2,6.5,6.25,7.8,9,cluster3
7,7.09,20.82,11.27,0.64,1.82,2.45,5.55,93.7,5.4,7.16,4.1,5.25,5.5,5.13,5.5,6,5.2,6.13,5.25,8.3,10,cluster3
8,1.82,13.73,6.82,9.27,4.18,4.82,6,61.7,3.1,7.74,3.4,4.75,4.75,4.75,5.83,5.75,5.2,4.75,4.38,6.6,11,cluster0
9,4.82,24.09,19.36,0.36,1.73,2.45,18.55,85,4.2,5.47,3.6,4.5,5,4,5.33,4.25,4.4,4.63,4.88,6.9,12,cluster1
10,2.73,15.36,8.36,4.18,1.82,3.09,4.36,93.78,4.44,6.79,3.2,4,4.5,2.25,5.33,5.75,6,4.88,5.63,7.85,14,cluster1
11,5.64,10.73,9.18,3.18,3.64,5.09,8.45,91.6,4.8,6.89,3.6,4,4.25,4.13,3.83,4.25,2.4,3.5,4.63,7.15,15,cluster1
12,5.73,17.91,16.36,0.82,2.09,2.64,12.27,139.2,4.7,6.84,3.8,5.5,6,4.38,5.83,6,3.2,4.38,5.25,7.05,16,cluster3
13,6.09,15.45,18.45,1.73,1.73,2.45,11.27,114.2,3.8,6.47,3.4,6,4,5.5,6.33,6,5.2,4.63,5.5,6.65,17,cluster3
14,11.82,25.09,13,0.18,1.45,2.36,7.64,180.78,7.89,9.32,4,6.5,6.25,6.5,7,6.5,3.8,6.75,5.75,9.8,18,cluster3
15,3.45,10.36,3,2.91,3.55,4.82,1.73,95,3,7.26,4.3,6.25,4,6.75,6.67,6.75,4.8,6,5.5,6.15,19,cluster3
16,4.18,17,7.82,4.91,3.18,3.55,6.18,94.7,4.6,6.63,3.9,5.5,4.75,5,5.67,5.25,4.4,4.13,6.13,6.05,20,cluster0
17,5.09,13.64,10.64,5.36,2.45,4.55,9.36,119.33,4.11,4.79,3.1,4.75,4,3.88,5.83,6.25,5.8,4.25,4,6.3,21,cluster1
18,7.91,10.64,10.18,0.82,2.55,4.18,8.27,127.5,3.8,7.95,4.2,6,3.25,5,4.83,5.5,3.8,4.5,5.25,6.5,23,cluster3
19,6.09,15.82,15.09,2.64,1.73,2.64,3.82,115,4,7.32,3.8,5,5,5.75,4.5,5,2,6.38,5,7.5,25,cluster3
20,6,16.73,7.64,2.36,2.27,2.91,7.27,88.8,3.2,7.42,3.9,5.25,5.5,5.25,5.5,5.25,2,6.38,5,8.75,26,cluster3
21,5.91,15.64,16.73,0.82,1.64,4.27,14.27,101.7,4.1,8.58,4.6,4,6,6.25,6.17,5.5,2.2,5.5,5,8.1,28,cluster3
22,7.18,12.64,16.55,2.73,3.91,4.55,13.55,155.33,5.11,7.16,3.7,6.5,5.5,5.13,5.83,6,5.6,5,5.13,7.85,29,cluster3
23,4.64,14.73,7.64,2.27,3.27,4.64,2.73,64,2.5,6,3.4,3.5,5.5,4.88,4.83,4.5,4.2,4.38,4.5,7.65,32,cluster1
24,5.91,12.18,15.64,3.09,3.36,4.73,14.64,105.67,3.56,5.47,3,5,4.75,2.63,5.67,5.5,6.2,4,5.38,6.75,34,cluster1
25,7.55,17.09,12.36,1.64,2,2.55,7.73,120.83,4.17,7.32,3.6,5,6,5.13,5.5,5.5,6.2,5.88,5.5,8.05,36,cluster3
26,10.09,20.55,17.18,1,2.27,3.27,13.91,92.6,4.5,6.26,3.7,5.25,3.5,4.5,4.83,4.5,3,4.5,5.13,6.4,37,cluster1
27,4.73,14.64,1.64,5.09,3.45,5.64,0.64,106,3,5.79,2.9,5.5,2.25,3.88,3.83,3.5,4,4.75,5.25,7.95,39,cluster0
28,5.36,13.45,13.18,4.91,5.73,5.82,10.64,109.33,4.89,6.42,3.6,4.75,4.75,5.75,5.67,5.75,6.2,4.63,4.75,6.15,41,cluster0
29,10.45,16,11.73,2.09,2.73,5.73,9.18,150.22,6,6.89,3.8,3.75,4.25,4.38,5,5,3.6,5.38,5.25,8.35,42,cluster1
30,4.27,18.91,8.45,5.64,4.18,5.09,6.45,93.3,2.7,7.74,3.2,4.5,5.25,4.38,3.67,5,6.2,4.13,4.5,6.7,43,cluster0
31,2.82,16.36,2.27,8.18,4,6.82,1.55,100.5,4.25,7.63,3.7,6,5,5.88,5.5,3.5,4.6,4.38,4.75,6.6,44,cluster0
```

Figura 3.2: Fichero obtenido en la agrupación automática.

En el presente estudio se propuso dos formas de agrupar estudiantes para mejorar el modelo de proceso educativo (Romero et al., 2016):

- **Manual**: se agrupan a los estudiantes usando solo la nota final obtenida en el curso, de forma que se ha detectado dos tipos de alumnos: los alumnos cuya nota final es menor a 5 (alumnos suspensos), y los alumnos cuya nota final es mayor o igual a 5 (alumnos aprobados).

- **Automática**: se agrupa a los estudiantes en base a la interacción que los estudiantes realizan durante el curso en la plataforma Moodle. En este caso las variables utilizadas para realizar el agrupamiento son variables relacionadas con el tiempo de trabajo, variables relacionadas con la procrastinación y variables relacionadas con la participación en foros. Estas variables tienen un valor determinado para cada uno de los alumnos que se estudian en este trabajo. En función de los valores que tengan asociados los sujetos a estas variables se les vinculará con uno de los tres grupos de nuestro estudio. Este agrupamiento se realizó con el algoritmo de agrupamiento EM[3] (Esperanza-Maximización) de la herramienta WEKA, (Witten et al., 2011) (ver figura 3.3).

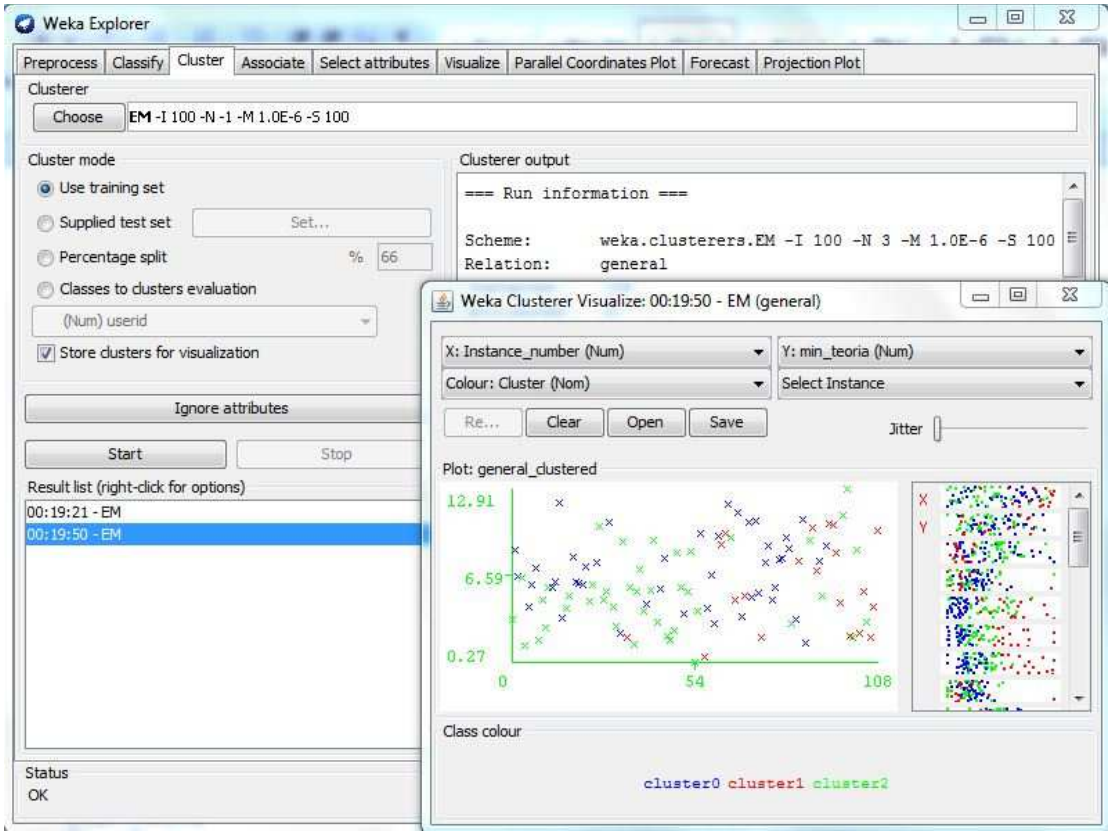


Figura 3.3: Interfaz de agrupamiento de WEKA.

En un último estudio (Bogarín et al., 2018b), se propuso la utilización de una codificación y agrupación de alto nivel (Fayyad et al., 1996) de los eventos de bajo nivel que captura Moodle durante la interacción del alumno con el curso. Concretamente, se propuso

[3] Se utilizó este algoritmo por ser un algoritmo de agrupamiento conocido que no requiere que el usuario especifique el número de grupos.

una taxonomía de cinco categorías principales: aprendiendo (LEARNING), planeando (PLANNING), ejecutando (EXECUTING), revisando (REVIEW) y aprendizaje colaborativo a través del foro (FORUM PEER LEARNING) (ver tabla 3.2). Los modelos de procesos generados consideran como "caso" al estudiante. El "evento" es la unión del campo acción más la codificación de alto nivel, por tanto, nuestros eventos son del tipo url view-LEARNING, quiz view-PLANNING, etc. Cada fila de nuestro registro de eventos es un evento (acción más codificación de alto nivel) que realiza un caso (estudiante) en una determinada fecha. Por tanto, la trazabilidad de cada caso son los diferentes eventos realizados por el estudiante. Por ejemplo, trazabilidad Estudiante A: <page view-LEARNING, quiz review-REVIEW, url view-LEARNING, quiz view-PLANNING….>

Tabla 3.2: Codificación de alto nivel para las acciones.

Acciones de Moodle de bajo nivel	**Codificación de alto nivel**
assign submit	EXECUTING
assign view	PLANNING
forum add discussion	FORUM PEER LEARNING
forum add post	FORUM PEER LEARNING
forum update post	FORUM PEER LEARNING
forum view discussion	FORUM PEER LEARNING
forum view forum	FORUM PEER LEARNING
page view	LEARNING
quiz attempt	EXECUTING
quiz close attempt	EXECUTING
quiz continue attempt	EXECUTING
quiz review	REVIEW
quiz view	PLANNING
quiz view summary	PLANNING
resource view	LEARNING
url view	LEARNING

Finalmente, todos los ficheros obtenidos para realizar los experimentos tienen que ser transformados a formato XES para que puedan ser procesados por la herramienta de minería de procesos ProM, utilizada en esta tesis. Esta herramienta de código abierto contiene una gran variedad de plug-ins para aplicar técnicas de minería de procesos.

3.3 Ejecución y comparación de algoritmos

En un primer estudio experimental (Romero et al., 2016) la metodología seguida se puede ver en la figura 3.4. Aplicamos el algoritmo HM sobre nuestros conjuntos de datos. Se utiliza HM porque además de ser el algoritmo más usado en entornos educativos se diseñó en una métrica basada en la frecuencia. Esto provoca que sea menos sensible al ruido y a registros de eventos que no están completos, características frecuentes en conjuntos de datos educativos extraídos de situaciones reales de aprendizaje. Las medidas de calidad que utilizamos fueron el ajuste y la complejidad de las redes obtenidas.

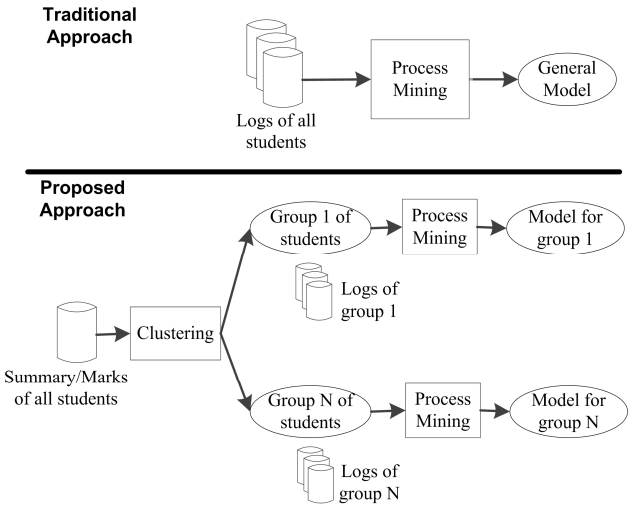


Figura 3.4: Nuestra propuesta VS investigación tradicional.

En un segundo estudio (Bogarín et al., 2018b) nuestros experimentos se centraron en los conjuntos de datos obtenidos del agrupamiento manual. En esta investigación incluimos varias novedades con respecto al anterior estudio (Romero et al., 2016). Por un lado, las pruebas se realizaron por temas para analizar más exhaustivamente el comportamiento de los estudiantes. Tuvimos que analizar el campo información de cada registro para vincularlos con los temas. Por otro lado, utilizamos una codificación de alto nivel con cinco etiquetas con el objetivo de conseguir modelos más comprensibles de

acuerdo con los supuestos de la teoría de SRL (Zimmerman, 1990). Se realizó un análisis comparativo de los modelos de procesos obtenidos con los algoritmos más utilizados en entornos educativos (Bogarín et al., 2018a): Alpha Miner (AM), HM, Evolutionary Tree Miner (ETM) e Inductive MIner (IM). Se utilizaron cuatro medidas de calidad para la evaluación de los modelos: ajuste, precisión, generalización y simplicidad (ver figura 3.5).

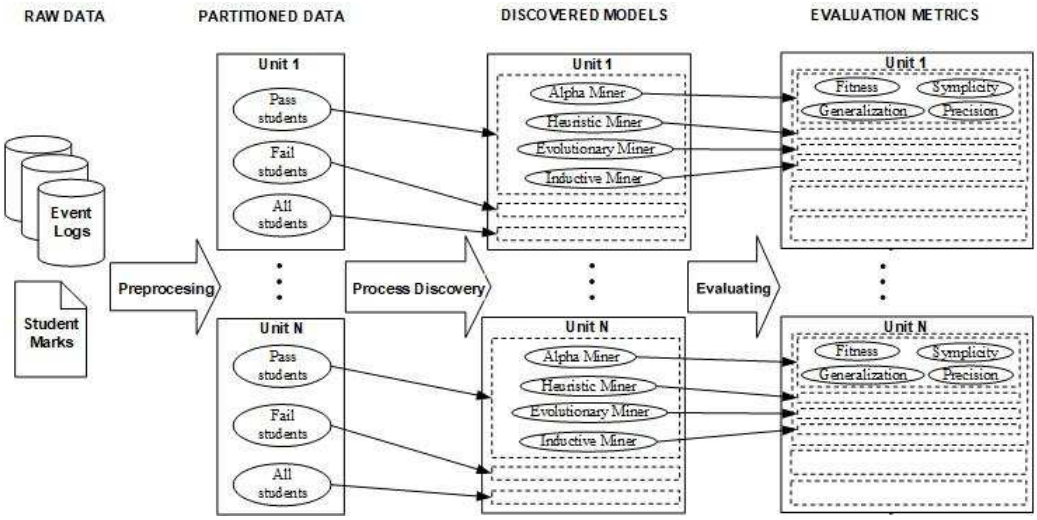


Figura 3.5: Procedimiento seguido para analizar EPM.

Todas estas medidas obtenidas en los modelos son importantes, pero no tiene sentido considerar la precisión, generalización y simplicidad si el ajuste no es bueno. Los algoritmos de descubrimiento de procesos normalmente consideran, como máximo, dos de las cuatro fuerzas principales de calidad porque cuando se optimiza la calidad de una, se pierde en otras. Alternativamente, utilizamos una nueva medida general propuesta por (Buijs et al., 2012), con el objetivo de balancear estas cuatro medidas conjuntamente. Se asignó un peso de 10 al ajuste, 5 a la precisión y 1 a la generalización y simplicidad (ver figura 3.6).

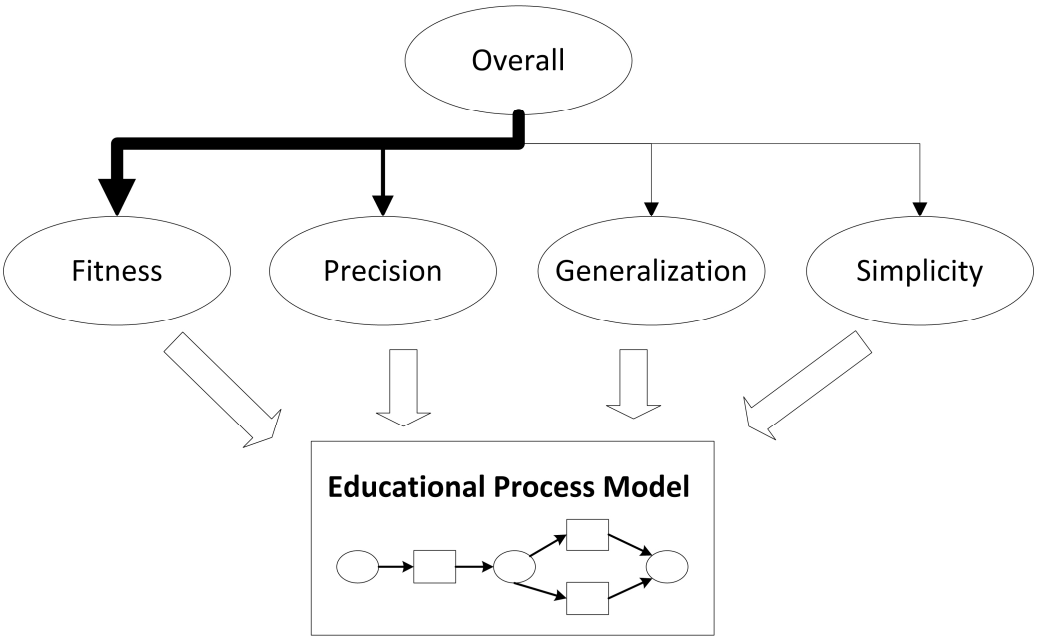


Figura 3.6: Métricas de calidad.

4

RESULTADOS

Tras la realización de los experimentos descritos en los diferentes artículos publicados en revistas con índice de impacto, capítulo de libro y congresos internacionales, se ha obtenido un buen número de resultados. A continuación, se detallan los principales, resultantes de estas publicaciones.

4.1 Experimento 1

En esta investigación (Romero et al., 2016) desarrollamos un tutorial de ProM con un estudio de caso en el que se realizaron varias pruebas. En la primera, se utilizaron todos los datos del registro de 84 estudiantes. En la segunda, se dividió el archivo de registro original en dos sub-conjuntos de datos: uno que contiene 68 estudiantes que aprobaron y otro con 16 estudiantes que suspendieron. En la tercera, se utilizó el algoritmo de agrupamiento EM para agrupar alumnos con similares características. En nuestro caso, se obtuvieron tres grupos con la siguiente distribución de los alumnos:

- **Grupo 0**: 23 estudiantes (22 aprueban y 1 suspenden)
- **Grupo 1:** 41 estudiantes (39 aprueban y 2 suspenden)
- **Grupo 2**: 20 estudiantes (13 suspenden y 7 aprueban)

Se aplicó minería de procesos sobre los diferentes conjuntos de datos con el fin de descubrir los modelos y flujos de trabajo de los estudiantes en cada uno de estos ficheros. Se obtuvieron los modelos utilizando HM, el algoritmo más usado en ámbitos educativos (Bogarín et al., 2018a). La característica más importante del HM es la robustez para el ruido y las excepciones. El HM está basado en la frecuencia de patrones, debido a que concentra su comportamiento principal en el registro de eventos. El modelo descubierto por el algoritmo de HM es una red heurística dibujada como un grafo cíclico dirigido, el cual representa el comportamiento más frecuente de los estudiantes en los conjunto de datos utilizados.

En los primeros resultados que obtuvimos (Bogarín et al., 2016), se observaba que la red heurística generada por los estudiantes suspensos (ver figura 4.1) era más pequeña que la red heurística de todos los estudiantes, y sobre todo, de los sujetos aprobados. En los grafos de los alumnos que suspenden observamos que había un significativo número de dependencias simples entre un conjunto de actividades (dos actividades o arcos que salen y entran en la misma actividad). Las actividades relacionadas con los *quiz* fueron las únicas que mostraban redes con una interacción mayor. Esto era coherente porque durante el curso esta actividad era obligatoria y puntuaba en la nota final. Con estos resultados, pudimos inferir que los estudiantes suspensos no utilizaban correctamente los recursos de la plataforma Moodle. Es fundamental que en una tarea existan ciertas relaciones entre determinadas acciones, lo que indica, que se han utilizado correctamente un repertorio de recursos concreto (necesario para poder aprobar cualquier asignatura). Esto si ocurría en los grafos de los estudiantes aprobados.

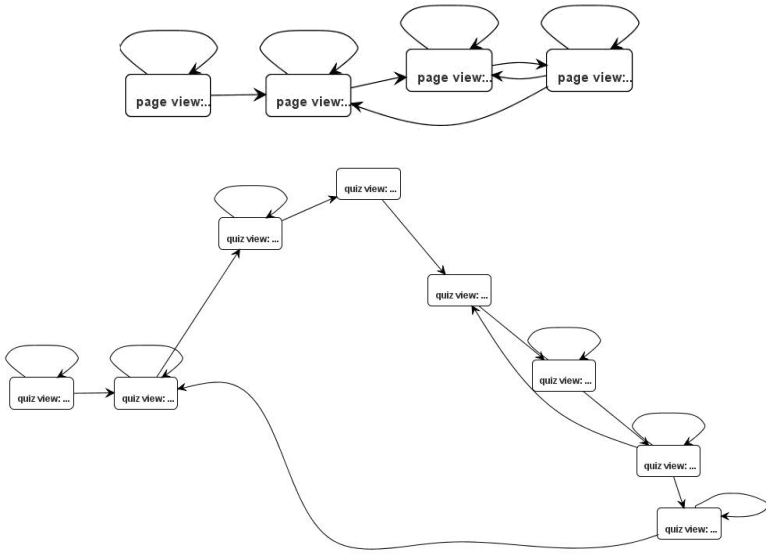


Figura 4.1: Red heurística de estudiantes suspensos.

Por otro lado, en el grafo de todos los alumnos, descubrimos nuevas redes que mostraban una mejor utilización de los recursos de Moodle porque contenían una mayor diversidad de acciones. Esto tiene sentido porque se mezclaban los estudiantes suspensos con los aprobados. Finalmente, se mostraron resultados gráficos obtenidos para los tres grupos generados en el agrupamiento automático. El grupo 0 es el que más subredes, número de nodos y enlaces contenía. Por tanto, es el grupo donde se ejecutaban una mayor variedad de actividades y con un cierto orden. Se observó correlación de estos gráficos con los anteriores, porque en este grupo es donde encontrábamos un mayor número de estudiantes aprobados.

La medida de calidad utilizada fue el ajuste, que indica la diferencia entre el comportamiento realmente observado en el registro y el comportamiento descrito por el modelo de proceso. El mayor valor de ajuste se obtuvo cuando usamos los datos de los estudiantes suspensos, donde 15 de los 16 estudiantes se ajustaban al modelo obtenido, es decir, el 93,75 %. Se pudo ver que estos modelos clusterizados, obtenidos usando agrupamiento manual y automático encajaban, mejor que el modelo general obtenido con todos los estudiantes (ver tabla 4.1).

Tabla 4.1: Ajuste de los modelos obtenidos.

Conjunto de Datos	**Ajuste**
Todos los estudiantes	0.8333
Estudiantes aprobados	0.9117
Estudiantes suspensos	0.9375
Estudiantes Cluster 0	0.9130
Estudiantes Cluster 1	0.9024
Estudiantes Cluster 2	0.9000

Por último, utilizamos dos medidas clásicas de la teoría de grafos (el número total de nodos y el número total de enlaces), con el fin de ver el nivel de complejidad de los modelos (ver tabla 4.2). El modelo más parsimonioso y por tanto más comprensible, se obtuvo con los estudiantes suspensos, seguido por el modelo de todos los estudiantes y el de los estudiantes del grupo 2. Por otro lado, los otros tres modelos eran cuantitativa y cualitativamente más complejos. Se cree que las razones fueron:

- En el conjunto de datos de todos los estudiantes, los estudiantes mostraron diferentes comportamientos y solo tenían algunas acciones en común porque había mezclados diferentes tipos de estudiantes (aprobados y suspensos).
- En el conjunto de datos de los estudiantes suspensos y del grupo 2, los estudiantes

mostraron solo algunos patrones de comportamiento común porque este tipo de estudiantes interactuaban menos con la plataforma Moodle.

- En el conjunto de datos de los estudiantes aprobados, grupo 0 y grupo 1, los estudiantes mostraron muchos más patrones de comportamiento comunes porque este tipo de estudiantes eran usuarios más activos en Moodle.

Tabla 4.2: Complejidad de los modelos obtenidos.

Conjunto de datos	N.Nodos	N.Enlaces
Todos los estudiantes	32	70
Estudiantes aprobados	113	244
Estudiantes suspensos	12	24
Estudiantes Cluster 0	61	121
Estudiantes Cluster 1	59	110
Estudiantes Cluster 2	38	84

Desde un punto de vista educativo y práctico esta información resulta útil para, entre otras, posibilitar a los profesores observar el proceso de aprendizaje del estudiante en situaciones no presenciales, detectar estudiantes en riesgo de suspender la asignatura o proporcionar *feedback* a los estudiantes.

4.2 Experimento 2

En el segundo estudio (Bogarín et al., 2018a) se realizó un análisis exhaustivo de la literatura publicada hasta la fecha acerca de minería de procesos en educación. Para ello comparamos EPM con otras áreas importantes relacionadas como Intentional Mining, SPM, y G-EDM. Se encontraron muchas semejanzas entre estas disciplinas, pero ninguna se centraba en el proceso de aprendizaje de una manera general, como EPM.

Asimismo, describimos los componentes que forman el marco de trabajo de EPM y los diferentes problemas que nos encontramos cuando tenemos que manipular un registro de eventos. Los agentes implicados en el proceso enseñanza-aprendizaje, lo entornos de aprendizaje virtuales, los registros de eventos y los modelos de procesos, resultaron ser los componentes principales de este nuevo marco de trabajo.

A continuación, explicamos los datos, herramientas, técnicas y modelos, usados en EPM. Al revisar la bibliografía observamos que el modelo de representación más usado en

EPM era Fuzzy net, seguido de la Petri net y Heuristic net. Por otro lado, el algoritmo de descubrimiento más utilizado era el HM y Fuzzy Miner (FM), y la técnica de conformidad más habitual el conformance checking.

Sin embargo, la aportación fundamental de esta investigación fue describir y agrupar los artículos seleccionados por dominios de aplicación educativos. Pudimos comprobar que los dominios más activos se resumían en investigaciones acerca de MOOC, AHS y LMS, CSCL y Professional Training.

En este sentido, los resultados obtenidos de este trabajo proporcionan una panorámica de los estudios realizados hasta la fecha con el objetivo de servir de guía a futuras investigaciones.

4.3 Experimento 3

En nuestra última investigación publicada (Bogarin et al., 2018b), propusimos la utilización de un nuevo algoritmo denominado Inductive Miner (Leemans et al., 2014), el cual no había sido aplicado previamente a datos educativos, con el objetivo de descubrir procesos de aprendizaje dentro de entornos de aprendizaje virtuales, conocidos en la literatura como VLEs. Asimismo, comparamos el rendimiento de IM con los algoritmos de minería de procesos educativos más utilizados en la literatura (HM, ETM y AM) (Bogarin et al., 2018a) para comprobar si se obtenían mejores modelos. Se utilizaron los datos de interacción de 101 estudiantes en una asignatura del grado de Psicología de la Universidad de Oviedo implementada en Moodle 2.0 durante el curso académico 2014-15. Utilizamos el fichero de eventos generado por esta plataforma educativa para realizar minería de procesos, descubrir modelos con los diferentes algoritmos y poder comparar sus medidas de ajuste, precisión, generalización y simplicidad, entendidas estas como:

- Ajuste: ¿es el modelo capaz de reproducir el comportamiento observado? Es decir, con esta medida podemos conocer si se pueden reproducir todas las trazas en el modelo obtenido.
- Precisión: ¿el modelo permite mucho más comportamiento del observado? Es decir, con esta medida nos muestra si el modelo permite una trazabilidad que no está entre las analizadas.
- Generalización: describe el sistema y no solo los datos. Aunque no se haya visto todo el comportamiento, se tiene que describir de una manera concisa. Es decir, esta medida podemos saber si se introduce comportamiento que no se ha observado.

- Simplicidad: El modelo es simple y fácil de entender. Se tiene que evitar los modelos de espagueti.

Los diferentes experimentos realizados en esta investigación se hicieron agrupando los datos del curso por temas, y se utilizó una codificación de alto nivel. En todas las pruebas realizadas con nuestro conjunto de datos, el algoritmo IM obtuvo las mejores métricas, sobre todo en el valor del ajuste que es la medida más relevante para el descubrimiento de modelos y el *overall* que engloba a todas las métricas (ver tabla 4.3). De hecho, no sirve de nada tener una buena medida de precisión o simplicidad, si no tenemos un buen ajuste. Además, cuando ponderamos con pesos las diferentes métricas según su relevancia, seguimos obteniendo la mejor medida general con el algoritmo propuesto.

Tabla 4.3: Comparación de los algoritmos respecto de la medida overall

Temas	**AM**	**HM**	**ETM**	**IM**
Tema 1	**0,676**	**0,666**	**0,793**	**0,797**
Tema 2	**0,666**	**0,618**	**0,752**	**0,781**
Tema 3	**0,583**	**0,493**	**0,675**	**0,712**
Tema 4	**0,452**	**0,597**	**0,715**	**0,747**
Tema 5	**0,582**	**0,533**	**0,649**	**0,659**
Tema 6	**0,577**	**0,621**	**0,742**	**0,793**
Tema 7	**0,612**	**0,664**	**0,724**	**0,773**
Tema 8	**0,724**	**0,732**	**0,750**	**0,796**
Tema 9	**0,516**	**0,510**	**0,744**	**0,784**
Tema 10	**0,685**	**0,700**	**0,827**	**0,856**
Tema 11	**0,553**	**0,563**	**0,735**	**0,778**

Para una de las unidades de conocimiento seleccionadas (tema 4) -se seleccionó porque es donde los estudiantes interactuaron en mayor medida con la plataforma en base al número de eventos-, observamos que el algoritmo que obtiene un mejor comportamiento en la medida *overall* es el IM, siendo incluso mejor cuando aplicamos clusterización (ver tabla 4.4). En concreto, se obtuvieron los mejores resultados en *el ajuste* y la *generalización*. En lo que respecta a *simplicidad*, los valores volvieron a ser los más altos, junto *al* ETM, por lo que los modelos obtenidos son fáciles de interpretar, evitando los

spaguetti[4]. Sin embargo, no se obtuvieron los mejores índices en *precisión* con el IM. En este sentido, los algoritmos que consiguieron mejores métricas de *precisión* en nuestro conjunto de datos son el ETM y el HM, con la diferencia de que el algoritmo ETM obtuvo métricas de *generalización* mejores.

Tabla 4.4: Comparación de los algoritmos en el tema 4 respecto de todas las métricas de calidad

Algoritmo	**Cluster**	**Fitness**	**Precision**	**Generalization**	**Simplicity**	**Overall**
Alpha Miner	Fail	0,765	0,197	0,422	0,636	0,570
Heuristic Miner	Fail	0,491	0,521	0,487	0,653	0,509
ET Miner	Fail	0,684	0,709	0,873	0,913	0,716
Inductive Miner	Fail	0,96	0,322	0,957	0,882	**0,768**
Alpha Miner	Pass	0,863	0,164	0,464	0,666	0,622
Heuristic Miner	Pass	0,526	0,707	0,603	0,732	0,596
ET Miner	Pass	0,712	0,691	0,841	0,901	0,725
Inductive Miner	Pass	0,959	0,315	0,962	0,882	**0,765**
Alpha Miner	All	0,581	0,198	0,414	0,466	0,452
Heuristic Miner	All	0,472	0,868	0,483	0,611	0,597
ET Miner	All	0,693	0,715	0,719	0,923	0,715
Inductive Miner	All	0,87	0,443	0,867	0,909	**0,747**

[4] Modelos muy complejos y difíciles de interpretar por el gran número de nodos y arcos que contienen.

Asimismo, se describió la utilidad de algunos de los modelos obtenidos por el IM en el tema 4. En este sentido, observamos que la primera actividad que realizaron los estudiantes suspensos (ver figura 4.2) en su ruta es *quiz attemp*, seguida de *quiz view summary* y *quiz view*. Todas ellas son actividades relacionadas con los *quiz*, una de las tareas obligatorias del curso, pero no la que daría comienzo a un proceso de aprendizaje tanto lógico como acorde al criterio del profesor o experto. En la parte central del modelo, realizaron actividades relacionadas con los foros como *forum view forum* y *forum add post*, la otra actividad obligatoria. A continuación, se observó que hay un paralelismo de las actividades *quiz review, quiz continue attemp* y *page view*, finalizando con la actividad de *url view*, punto de partida lógico sugerido por el profesor para un aprendizaje óptimo. Además, hay cierto tipo de acciones que sí realizaron los estudiantes aprobados y no los suspensos, como son *forum update* post y *forum view discussion*. Este comportamiento resulta especialmente interesante teniendo en cuenta que la actividad en los foros se ha relacionado, en estudios previos, con el rendimiento del estudiante en los LMS (Romero et al., 2013). Por otro lado, observando la codificación de alto nivel, los modelos nos llevaron a concluir que los estudiantes suspensos no siguieron la ruta de aprendizaje sugerida por el profesor y promovida por la teoría de SRL que conducen a resultados de aprendizaje de calidad. Basándonos en las asunciones del SRL, comenzar el proceso de aprendizaje realizando acciones de EXECUTING sin realizar previamente nada de LEARNING o PLANNING es un camino que conduce al estudiante al fracaso, como así finalmente se muestran los datos de este estudio.

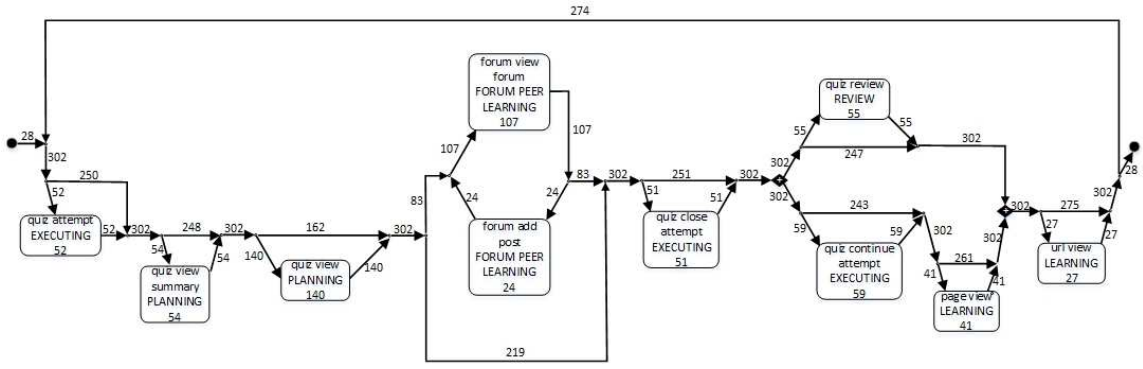


Figura 4.2: Modelo obtenido en el tema 4 para los estudiantes suspensos.

En cuanto a la ruta seguida por los estudiantes aprobados (ver figura 4.3), se pudo ver que comenzaron visitando el *forum view discussion* y, a continuación, siguen hasta tres rutas diferentes. Una de las rutas es realizar LEARNING a través de la actividad *url view*. Otro posible camino sería realizar en paralelo una serie de actividades relacionadas con los foros *forum view forum, forum add post* y *forum update post*. En la última ruta, vimos que se

ejecutaron actividades relacionadas con los *quiz* donde realizan EXECUTING, PLANNING y LEARNING. También pudimos observar que en las diferentes rutas hay una homogeneidad entre las actividades y que se terminaron con la actividad *quiz close attempt* y *quiz review.* Observando la codificación de alto nivel de los aprobados pudimos deducir que, aunque no siguieron exactamente las sugerencias del profesor, trazaron un proceso que en mayor o menor medida se acercaba a ellas, y finalmente les condujo al éxito. Estos resultados están en línea con los de Lust et al. (2013a, 2013b), que concluyeron que solo una minoría de estudiantes regulaba su comportamiento de acuerdo con los requisitos del curso.

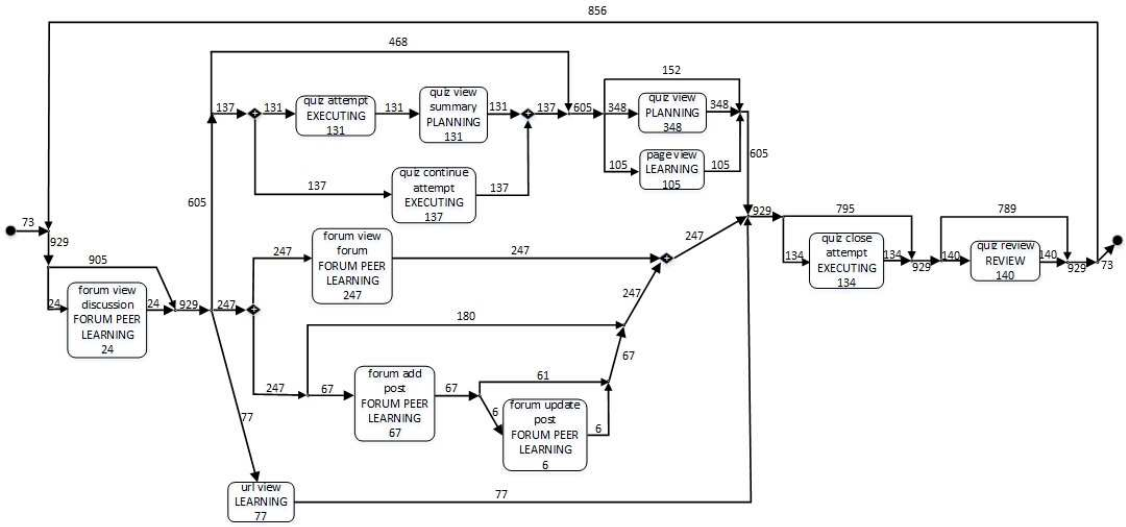


Figura 4.3: Modelo obtenido en el tema 4 para los estudiantes aprobados.

5

CONCLUSIONES

En esta tesis se ha propuesto la aplicación de algoritmos de minería de procesos sobre datos de los registros de interacción de los estudiantes (ficheros logs) recogidos por el entorno de educación en línea Moodle. Nuestro objetivo era descubrir modelos visuales, fiables y comprensibles sobre las rutas seguidas por los estudiantes durante su proceso de aprendizaje de forma que puedan ser de utilidad a profesores e investigadores, para comprender mejor tanto ese proceso como su resultado.

En nuestra primera hipótesis planteamos llevar a cabo el estado del arte sobre EPM para conocer cuáles eran los algoritmos y herramientas más utilizadas y con mejores resultados. Para confirmar esta hipótesis en nuestros primeros experimentos y, tras revisar la literatura previa sobre aplicación de minería de procesos en educación, se observó que los algoritmos AM, HM y FM eran los algoritmos más utilizados para descubrir modelos de aprendizaje (Bogarín et al., 2018a), concretamente HM era el que mejores resultados mostraba. Posteriormente, descubrimos que con el nuevo algoritmo IM se podían obtener mejores resultados (Bogarín et al., 2018b) que con estos algoritmos tradicionales, incluido HM. Desde un punto de vista práctico, estos resultados obtenidos con IM pueden suponer una aportación fundamental para detectar estudiantes que corren el riesgo de suspender un curso o asignatura o desviaciones de los estudiantes en las rutas de aprendizaje que en cada caso funcionen como criterio. Para intervenir en este sentido, se podrían diseñar acciones preventivas en entornos de aprendizaje hipermedia (Brusilovsky & Millán, 2007) o sistemas de detección temprana.

Asimismo, es importante resaltar que EPM no es interesante únicamente a nivel retrospectivo, sino que también es relevante en el presente, pudiéndose explotarse a través de recomendaciones en tiempo real, y por supuesto, para intervenir en la conducta futura, por ejemplo, a través de predicciones de comportamientos disregulados futuros (van der Aalst et al., 2011).

Obtener modelos válidos e interpretables fue el objetivo central de esta tesis, en este sentido, hemos conseguido que los modelos visuales sean comprensibles para los docentes y los estudiantes, lo que hace que estos resultados sean esenciales para monitorizar el proceso de aprendizaje y proporcionar feedback. Las metodologías empleadas para la consecución de este objetivo se encuentran en Bogarín et al. (2016) y Bogarín et al. (2018b). Con el objetivo de comprobar la tercera hipótesis planteada en esta tesis, propusimos, con éxito, diferentes tipos de agrupamientos (manual y automático) para mejorar los modelos de EPM y, al mismo tiempo, optimizar el rendimiento (métricas) y la comprensibilidad (tamaño) de los modelos. Sin embargo, en Bogarín et al. (2018b) se realizaron las pruebas por temas o unidades de conocimiento, y conseguimos analizar en mayor profundidad el comportamiento de los estudiantes; de esta manera hemos obtenido modelos más específicos que son, además, más certeros. Para satisfacer la segunda hipótesis de esta tesis y codificar los ficheros de datos proporcionados por Moodle, en lugar de los eventos de bajo nivel, utilizamos una codificación de alto nivel con cinco etiquetas y obtuvimos un nivel de abstracción mayor y modelos más comprensibles y sencillos desde el punto de vista de los supuestos de la teoría de SRL.

Los modelos que hemos obtenido en nuestros estudios resultaron de gran utilidad ya que el profesor obtiene una información más completa de cómo está trabajando el estudiante durante el curso, pudiendo detectar en una fase más temprana rutas que llevan a futuras dificultades o un mal rendimiento.

Por otro lado, con la finalidad de verificar la cuarta hipótesis de la tesis comparamos los diferentes algoritmos de descubrimiento de modelos de procesos utilizando varias medidas de calidad. La utilización de diferentes métricas de evaluación de los modelos obtenidos (Bogarín et al., 2018b), nos sirvió para contrastar de manera empírica tres importantes conclusiones:

1. Con el algoritmo IM obtenemos los mejores resultados en la medida del ajuste. Esto es importante porque no tiene sentido considerar otras medidas si el ajuste no es bueno (Buijs et al., 2012).
2. Los resultados obtenidos en el balanceo de las métricas de calidad (overall) son mejores en el IM que en otros algoritmos tradicionales de EPM.

3. Los resultados obtenidos en las métricas, analizadas en conjunto o individualmente, son aún mejores en los conjuntos de datos que estaban agrupados, como se vio con datos educativos en (Bogarín et al., 2016).

Por tanto, la utilización del nuevo algoritmo IM para descubrir modelos de aprendizaje abre un nuevo campo en la investigación, el desarrollo y la comprensión de PM aplicado a entornos educativos.

5.1 Futuras mejoras

EPM permite una mejor comprensión del proceso educativo subyacente en los registros de eventos de los VLEs, pero como toda disciplina nueva y emergente, se enfrenta a muchos desafíos y perspectivas de futuro. Una ambiciosa futura mejora sería desarrollar un plugin para Moodle que permita obtener los modelos en tiempo real. Actualmente, Moodle no proporciona ninguna herramienta de descubrimiento de los modelos de rutas seguidos por los estudiantes cuando interactúan con esta plataforma. Con este tipo de herramienta acercaríamos estas potentes técnicas de minería de procesos a los diferentes agentes involucrados en el proceso de aprendizaje y se rentabilizaría la gran cantidad de datos en bruto generados para transformarlos en una estructura comprensible para su uso posterior.

Asimismo, con el objetivo de reafirmar y generalizar los buenos resultados obtenidos con el algoritmo IM con datos de un curso Moodle, se deberían replicar las pruebas en cursos de diferentes áreas de conocimiento, para poder generalizar los resultados. Además, sería interesante probar este algoritmo con datos de diferentes plataformas de educación en línea como los emergentes MOOCs.

Finalmente, algunas otras cuestiones que creemos que serán especialmente relevantes en el futuro cercano del EPM:

- Desarrollar herramientas de EPM específicas para educación (Barreiros et al., 2014) y así poder acercar la minería de procesos a los expertos y especialistas en educación e investigadores, que, en general, no poseen los conocimientos técnicos necesarios para utilizar las herramientas de software de EPM como ProM.
- Añadir información semántica para mejorar EPM (Okoye et al., 2016; Sedrakyan et al., 2016). Los conceptos semánticos se pueden superponer a las acciones realizadas por los estudiantes con el objetivo de proporcionar un análisis más conceptual de los procesos y proporcionar respuestas del mundo real que están más cerca de la comprensión humana.

- Aplicar recomendación en EPM (Chen et al., 2015; Bannert et al., 2014; Khodabandelou et al., 2013; Reimann et al., 2014; Wang & Zaïane, 2015). La información descubierta en los PM debe ser, no solo comprensible, sino también útil para la toma de decisiones de los usuarios finales, por lo que los resultados obtenidos pueden ser presentados en forma de recomendaciones para los profesores o investigadores de EPM.
- Aplicar EPM en otros sub-dominios educativos emergentes como juegos, dispositivos móviles y entornos de aprendizaje ubicuos (Porouhan & Premchaiswadi, 2017). Actualmente las técnicas de EPM ya han comenzado a aplicarse en juegos (Vermeulen et al., 2016) para analizar el comportamiento de los estudiantes. Asimismo, los dispositivos móviles digitales como tabletas, PDA (Personal Digital Assistant) y teléfonos inteligentes, también se utilizan, cada vez con mayor frecuencia, con fines educativos.
- Hacer públicos y libres conjuntos de datos de eventos (ficheros logs) procedentes de diferentes entornos educativos. Creemos que es esencial que más conjuntos de datos estén disponibles libremente para que la investigación sobre EPM avance. De hecho, solo hemos encontrado un conjunto de datos de EPM específico libre[5]. Por lo tanto, uno de los siguientes pasos sería promover el intercambio de colecciones de datos con el objetivo de que se pueden analizar desde múltiples perspectivas, con diversos métodos y herramientas.

5.2 Contribuciones científicas

Los artículos publicados en revista con índice de impacto, capítulo de libro, congresos internacionales y revistas nacionales que se presentan en la parte II de publicaciones de esta tesis se relacionan a continuación:

[5] https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A+A+Learning+Analytics+Data+Set

Revistas Indexadas

Artículo 1: Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1), 1-17.

Artículo 2: Bogarín, A., Cerezo, R., & Romero, C. (2018). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). *Psicothema*, *30*(3), 322-329.

Capítulo de Libro

Artículo 3: Romero, C., Cerezo, R., Bogarín, A., & Sánchez□Santillán, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. In: *ElAtia, S., Ipperciel, D., & Zaïane, O.R. (Eds.). Data Mining and Learning Analytics: Applications in Educational Research*. 3-28.

Congresos Internacionales

Artículo 4: Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In: *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. Indianápolis, USA. 11-15.

Artículo 5: Bogarín, A., Romero, C., & Cerezo, R. (2015). Discovering Students' Navigation Paths in Moodle. In: *Educational Data Mining (EDM)*. Madrid, Spain. 556-557.

Revistas Nacionales

Artículo 6: Bogarín, A., Romero, C. y Cerezo, Rebeca. (2015). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. In: *EDMETIC*, *5*(1), 73-92.

Figura 5.1: Publicaciones.

REFERENCIAS BIBLIOGRÁFICAS

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In: *Data Engineering. Proceedings of the Eleventh International Conference*. Taipei, Taiwan. 3-14.

Anuwatvisit, S., Tunggkasthan, A., & Premchaiswadi, W. (2012). Bootleneck mining and petri net simulation in education situations. In: *Conference on ICT and Knowledge Engineering*. Bangkok, Thailand. 244-251.

Ariouat, H., Cairns, A.H., Barkaoui, K., Akoka, J., & Khelifa, N. (2016). A Two-Step Clustering Approach for Improving Educational Process Model Discovery. In: *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 25th International Conference*. Paris, France. 38-43.

Ayutaya, N. S. N., Palungsuntikul, P., & Premchaiswadi, W. (2012). Heuristic mining: Adaptive process simplification in education. In: *International Conference on ICT and Knowledge Engineering*. Bangkok, Thailand. 221-227.

Azevedo, R., Behnagh, R., Duffy, M., Harley, J. M., & Trevors G. J. (2012). Metacognition and self-regulated learning in student-centered learning environments. In: *Theoretical foundations of student-center learning environments, 49*(2), 171–197.

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. In: *Metacognition and learning, 9*(2), 161-185.

Barreiros, B.V., Lama, M., Mucientes, M., & Vidal, J.C. (2014). Softlearn: A process mining platform for the discovery of learning paths. In: *14th International Conference on Advanced Learning Technologies*. Athens, Greece. 373-375.

Bergenthum, R., Desel, J., Harrer, A., & Mauser, S. (2008). Learnflow Mining. In: *Die 6. e-Learning Fachtagung Informatik.* Lübeck, Germany. 269-280.

Bergenthum, R., Desel, J., Harrer, A., & Mauser, S. (2012). Modeling and mining of learnflows. In: *Jensen, K., Donatelli, S., & Kleijn, J. (Eds.). Transactions on Petri Nets and Other Models of Concurrency V.* 22-50.

Bogarín, A., Cerezo, R., & Romero, C. (2018a). A survey on educational process mining. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(1), 1-17.

Bogarín, A., Cerezo, R., & Romero, C. (2018b). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). In: *Psicothema, 30*(3), 322-329.

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In: *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. Indianapolis, USA. 170-181

Boiarsky, C. (1984). A model for analyzing revision. In: *Journal of Advanced Composition, 5*, 65-78.

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In: *Brusilovsky, P., Kobsa, A., & Nejdl, W. (Eds.). The Adaptive Web: Methods and Strategies of Web Personalizatio*. 3-53.

Buijs, J. C., Van Dongen, B. F., & van der Aalst, W. M. (2012). On the role of fitness, precision, generalization and simplicity in process discovery. In: *Proceedings of the OTM Confederated International Conferences" On the Move to Meaningful Internet Systems".* Rome, Italy. 305-322.

Cairns, A.H., Gueni, B., Assu, J., Joubert, C., & Khelifa, N. (2015a). Process Mining in the Education Domain. In: *International Journal on Advances in Intelligent Systems, 8*(1), 219-232.

Cairns, A.H., Gueni, B., Assu, J., Joubert, C., & Khelifa, N. (2015b). Analyzing and Improving Educational Process Models using Process Mining Techniques. In: *The Fifth International Conference on Advances in Information Mining and Management*. Brussels, Belgium. 17-22.

Cairns, A.H., Gueni, B., Fhima, M., Cairns, A., David, S., & Khelifa, N. (2014a) Towards Custom-Designed Professional Training Contents and Curriculums through Educational Process Mining. In: *The Fourth International Conference on Advances in Information Mining and Management*. Paris, France. 53-58.

Cairns, A.H., Ondo, J.A., Gueni, B., Fhima, M., Schwarcfeld, M., Joubert, C., & Khelifa, N. (2014b). Using Semantic Lifting for Improving Educational Process Models Discovery and Analysis. In: *Fourth international symposium on data-driven process discovery and analysis* (*SIMPDA)*. Milan, Italy. 150-161.

Chen, J., Zhang, Y., Sun, J., Chen, Y., Lin, F., & Jin, Q. (2015). Personalized Micro-Learning Support Based on Process Mining. In: *7th International Conference on Information Technology in Medicine and Education (ITME)*. Huangshan, Anhui, China. 511-515.

Doleck, T., Jarrell, A., Poitras, E.G., Chaouachi, M., & Lajoie, S.P. (2016). Examining Diagnosis Paths: A Process Mining Approach. In: *Computational Intelligence & Communication Technology (CICT), Second International Conference*. Ghaziabad, India. 663-667.

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. (2015). Clustering algorithms applied in educational data mining. In: *International Journal of Information and Electronics Engineering, 5(*2), 112.

Emond, B., & Buffett, S. (2015). Analyzing Student Inquiry Data Using Process Discovery and Sequence Classification. In: *International Educational Data Mining Society*. Madrid, Spain. 412-416.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. In: *Communications of the ACM, 39*(11), 27-34.

Fernández-Gallego, B., Lama, M., Vidal, J. C., & Mucientes, M. (2013). Learning analytics framework for educational virtual worlds. In: *Procedia Computer Science, 25,* 443-447.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. In: *AI magazine*, *13*(3), 57-70.

Günther, C. W., & van der Aalst, W. M. (2007). Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In: *International Conference on Business Process Management*. Brisbane, Australia. 328-343.

Howard, L., Johnson, J., & Neitzel, C. (2010). Examining learner control in a structured inquiry cycle using process mining. In: *Educational Data Mining*. Pittsburgh, USA. 71-80.

Jeong, H., Biswas, G., Johnson, J., & Howard, L. (2010). Analysis of productive learning behaviors in a structured inquiry cycle using hidden markov models. In: *Educational Data Mining*. Pittsburgh, USA. 81-90.

Khodabandelou, G., Hug, C., Deneckere, R., & Salinesi, C. (2013). Process mining versus intention mining. In: *Enterprise, Business-Process and Information Systems Modeling*. Valencia, Spain. 466-480.

Leemans, S. J. J., Fahland, D., & van der Aalst, W. M. P. (2014). Process and Deviation Exploration with Inductive Visual Miner. In: *Business Process Management Demo Sessions (BPMD)*. Eindhoven, The Netherlands. 46-50.

Lust, G., Elen, J., & Clarebout, G. (2013a). Regulation of tool-use within a blended course: student differences and performance effects. In: *Computers & Education, 60*(1), 385-395.

Lust, G., Elen, J., & Clarebout, G. (2013b). Measuring students' strategy-use within a CMS supported course through students' tool-use patterns. In: *15th biennial conference EARLI.* Munich, Germany. 571-572.

Mekhala. (2015). A review paper on Process Mining. In: *International Journal of Engineering Research and Technology*, *1*(4), 11-17.

Mukala, P., Buijs, J., Leemans, M., & van der Aalst, W. (2015b). Learning Analytics on Coursera Event Data: A Process Mining Approach. In: *Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis*. Vienna, Austria. 18-32.

Mukala, P., Buijs, J., & van der Aalst, W.M.P. (2015a). Uncovering learning patterns in a MOOC through conformance alignments. In: *Tech. rep., Eindhoven University of Technology, BPM Center Report BPM*, *1509*.

Nesbit, J.C., Zhou, M., Xu, Y., & Winne, P. (2007). Advancing log analysis of student interactions with cognitive tools. In: *12th biennial conference of the european association for research on learning and insruction (EARLI)*. Budapest, Hungary. 2-20.

Okoye, K., Tawil, A.R.H., Naeem, U., & Lamine, E. (2016). Discovery and Enhancement of Learning Model Analysis through Semantic Process Mining. In: *International Journal of Computer Information Systems and Industrial Management Applications, 8,* 93-114.

Pechenizkiy, M., Trcka, N., Vasilyeva, E., van der Aalst, W., & De Bra, P. (2009). Process mining online assessment data. In: *Educational Data Mining*. Córdoba, Spain. 279-288.

Perez-Rodriguez, R., Caeiro-Rodriguez, M., & Anido-Rifon, L. (2009). Enabling process-based collaboration in Moodle by using aspectual services. In: *Ninth IEEE International Conference on Advanced Learning Technologies*. Riga, Latvia. 301-302.

Poncin, W., Serebrenik, A., & van den Brand, M. (2011a). Process mining software repositories. In: *15th European Conference on Software Maintenance and Reengineering (CSMR)*. Burlington, USA. 5-14.

Poncin, W., Serebrenik, A., & van den Brand, M. (2011b). Mining student capstone projects with FRASR and ProM. In: *Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion*. Portland, USA. 87-96.

Porouhan, P., & Premchaiswadi, W. (2017). Process Mining and Learners' Behavior Analytics in a Collaborative and Web-Based Multi-Tabletop Environment. In: *International Journal of Online Pedagogy and Course Design (IJOPCD), 7*(3), 29-53.

Reimann, P., Frerejean, J., & Thompson, K. (2009). Using process mining to identify models of group decision making in chat data. In: *Proceedings of the 9th international conference on Computer supported collaborative learning*. Rhodes, Greece. 98-107.

Reimann, P., Markauskaite, L., & Bannert, M. (2014). e-Research and learning theory: What do sequence and process mining methods contribute? In: *British Journal of Educational Technology, 45*(3), 528-540.

Romero, C., Cerezo R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: a tutorial and case study using moodle data sets. In: *ElAtia, S., Ipperciel, D., & Zaïane, O.R. (Eds.). Data Mining and Learning Analytics: Applications in Educational Research*. 1-28.

Romero, C., Lopez, M.I., Luna, J.M., & Ventura, S. (2013). Predicting students' final performance from participation in online discussion forums. In: *Computers&Education, 68*, 458–472.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. In: *Expert systems with applications, 33*(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. In: *IEEE Transactions on Systems, Man, and Cybernetics, 40*(6), 601-618.

Romero, C., & Ventura, S. (2017). Educational data science in massive open online courses. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(1), 1-12.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. In: *Computers & Education, 51*(1), 368-384.

Rozinat, A., & van der Aalst, W.M. (2005). Conformance testing: Measuring the fit and appropriateness of event logs and process models. In: *International Conference on Business Process Management.* Nancy, France. 163-176.

Schoonenboom, J., Levene, M., Heller, J., Keenoy, K., & Turcansyi, M. (2007). *Trails in education: Technologies that support navigational learning*. First Edition. SensePublishers.

Schoor, C., & Bannert, M. (2012). Exploring regulatory processes during a computer-supported collaborative learning task using process mining. In: *Computers in Human Behavior, 28*(4), 1321-1331.

Schulte, J., Fernandez de Mendonca, P., Martinez-Maldonado, R., & Buckingham Shum, S. (2017). Large scale predictive process mining and analytics of university degree course data. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. Vancouver, Canada. 538-539.

Sedrakyan, G., De Weerdt, J., & Snoeck, M. (2016). Process-mining enabled feedback: "Tell me what I did wrong" vs. "tell me how to do it right". In: *Computers in Human Behavior, 57,* 352-376.

Siemens, G., & d Baker, R.S. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: *Proceedings of the 2nd international conference on learning analytics and knowledge*. Vancouver, Canada. 252-254.

Southavilay, V., Yacef, K., & Callvo, R.A. (2010). Process mining to support students' collaborative writing. In: *Educational Data Mining*. Pittsburgh, USA. 257-266.

Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills. In: *Csapó, B., & Funke, J. (Eds.). The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*. 193-209.

Trcka, N., & Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. In: *Ninth International Conference on Intelligent Systems Design and Applications*. Pisa, Italy. 1114-1119.

Trcka, N., Pechenizkiy, M., & van der Aalst, W.M.P. (2011). Process mining from educational data. In: *Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.d. (Eds.). Handbook of educational data mining*. 123-142.

Vahdat, M., Oneto, L., Anguita, D., Funk, M., & Rauterberg, M. (2015). A Learning Analytics Approach to Correlate the Academic Achievements of Students with Interaction Data from an Educational Simulator. In: *Design for Teaching and Learning in a Networked World*. Toledo, Spain. 352-366.

Van der Aalst, W.M. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. First Edition. Springer.

Van der Aalst, W.M. (2016). *Process mining: data science in action*. Second Edition. Springer.

Van der Aalst, W.M., Guo, S., & Gorissen, P. (2013). Comparative process mining in education: An approach based on process cubes. In: International Symposium on Data-Driven Process Discovery and Analysis. Riva del Garda, Italy: 110-134.

Van der Aalst, W.M., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. In: *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1128-1142.

Van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., & van der Aalst, W.M. (2005). The ProM framework: A new era in process mining tool support. In: *International Conference on Application and Theory of Petri Nets*. Miami, USA. 444-454.

Vellido, A., Castro, F., & Nebot, A. (2011). Clustering Educational Data. In: *Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R.S.J.d. (Eds.). Handbook of Educational Data Mining*. 75-92.

Vermeulen, M., Mandran, N., & Labat, J.M. (2016). Chronicle of a scenario graph: from expected to observed learning path. In: *11th European Conference on Technology Enhanced Learning*. Lyon, France. 321-330.

Vidal, J.C., Lama, M., & Bugarín, A. (2012). Petri net-based engine for adaptive learning. In: *Expert Systems with Applications, 39*(17), 12799-12813.

Vidal, J.C., Vázquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. In: *Knowledge-Based Systems, 100*, 160-174.

Wang, R., & Zaïane, O.R. (2015). Discovering Process in Curriculum Data to Provide Recommendation. In: *Educational Data Mining.* Madrid, Spain. 580-581.

Washio, T., & Motoda, H. (2003). State of the art of graph-based data mining. In: *Acm Sigkdd Explorations Newsletter, 5*(1), 59-68.

Weijters, A.J.M.M., van der Aalst, W.M., & de Medeiros, A.A. (2006). Process mining with the heuristics miner-algorithm. In: *Tech. rep., Technische Universiteit Eindhoven, 166,* 1-34.

Witten, I.H., Eibe, F., & Hall, M.A. (2011). *Data Mining, Practical Machine Learning Tools and Techniques*. Third Edition. Morgan Kaufman Publishers.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. In: *Educational psychologist, 25*(1), 3-17.

Parte II. Publicaciones

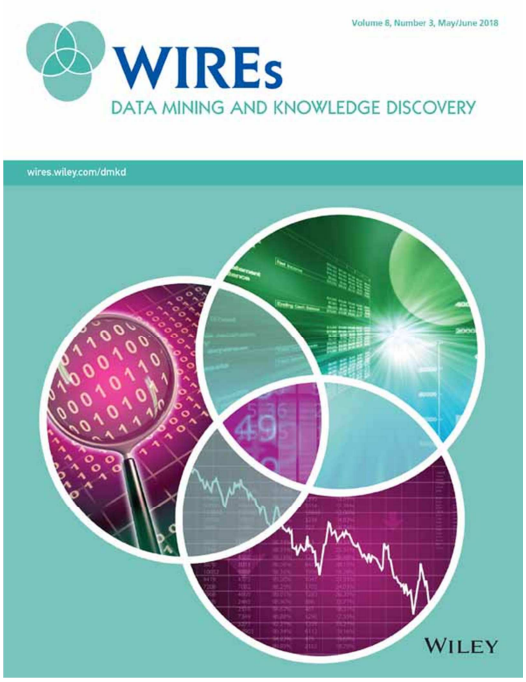
ARTÍCULO 1

Referencia:

- Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1), 1-17.

Medidas de Calidad Científica:

- Índice de impacto JCR (2017): 1.939
- Área: COMPUTER SCIENCE, THEORY & METHODS
- Cuartil en WOS: Q2
- Posición relativa dentro del área: 29 de 103
- Número total de citas en Web of Science: 4
- Número total de citas en GoogleShoolar: 6



Overview

A survey on educational process mining

Alejandro Bogarín,[1] Rebeca Cerezo[2] and Cristóbal Romero[1]*

Educational process mining (EPM) is an emerging field in educational data mining (EDM) aiming to make unexpressed knowledge explicit and to facilitate better understanding of the educational process. EPM uses log data gathered specifically from educational environments in order to discover, analyze, and provide a visual representation of the complete educational process. This paper introduces EPM and elaborates on some of the potential of this technology in the educational domain. It also describes some other relevant, related areas such as intentional mining, sequential pattern mining and graph mining. It highlights the components of an EPM framework and it describes the different challenges when handling event logs and other generic issues. It describes the data, tools, techniques and models used in EPM. In addition, the main work in this area is described and grouped by educational application domains. © 2017 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2018, 8:e1230. doi: 10.1002/widm.1230

INTRODUCTION

Nowadays, with the development and increasing popularity of technology supported learning environments, information systems enable us to capture all student events/actions/activities at different levels of granularity, from low level events such as keystrokes, mouse gestures and clicks, to higher-level events such as students' learning activities.[1] These systems have tracking and logging capabilities to gather different types of temporal data such as click streams, chat logs, document edit histories (e.g., wikis, etherpads), motion tracking (e.g., eye-tracking, Microsoft Kinect), learning resource usage logs, and various interaction logs (e.g., with intelligent tutoring systems). Process mining (PM) can use these so-called logs, audit trails, log files or traces in order to discover, monitor and improve educational processes. PM provides a bridge between data mining (DM) and process modeling and analysis. PM as a sub-discipline of DM adds the process-oriented view of the DM procedure.[2] DM has been widely applied successfully to find interesting patterns from data gathered in educational environments.[3] However, educational data mining (EDM) techniques focus on data dependencies or simple patterns and do not provide a visual representation of the overall learning process. EDM does not focus on the process as a whole.[4] Classical DM techniques such as classification, clustering, regression, association rule mining (ASR) and sequence mining are of little use for control-flow discovery and are not process-centric. To allow for these types of general analysis, in which the process rather than the result plays the central role, a new method of data-mining research, called PM, has been proposed.[1] More specifically, educational process mining (EPM) is the application of PM to raw educational data.[5] Both EDM and EPM apply specific algorithms to data in order to uncover hidden patterns and relationships. But EDM techniques are not process-centric and do not focus on event data.[2] For EDM techniques the rows (instances) and columns (variables) of a typical data file do not have any meaning. On the other hand, EPM is process-centric thereby making the unknown (or partially known) processes explicit; it assumes a different type of data: events. Each event belongs to a single process instance (also called a case), and events are related to activities. EPM is interested in end-to-end

*Correspondence to: cromero@uco.es

[1]Department of Computer Science, University of Cordoba, Cordoba, Spain

[2]Department of Psychology, University of Oviedo, Oviedo, Spain

Conflict of interest: The authors have declared no conflicts of interest for this article.

processes rather than local patterns. End-to-end process models and concurrency are essential for PM.[6] Furthermore, other suitable approaches with the same scope such as process discovery, conformance checking, and bottleneck analysis, are not addressed by traditional EDM techniques.

On this basis, this paper is arranged as follows: *Background and Related Areas* in second section. *Framework and Concepts* introduces the EPM framework. *Data and Tools* describes the type of data, formats and tools commonly used in EPM. *Techniques* section, outlines the most commonly used techniques. *Application Domain* describes the different application domains in education and the most relevant work. Finally, *Conclusions and Further Research* are outlined.

BACKGROUND AND RELATED AREAS

PM is a relatively new technology emerging from the business community.[2] It focuses on the development of techniques aimed at extracting process-related knowledge from event logs. It uses event logs recorded by information systems in order to discover, monitor and improve processes in different domains as well as to check process conformance, detect bottlenecks, and predict execution problems. PM is also known as Workflow Mining (WM). Most work in PM has concentrated on (business) workflow systems and the discovery of Petri nets representations of workflows.[7] The methods described by these labels take information from event logs as input and produce process models that describe the information in the event logs in a comprehensive manner.[8] The process-oriented view helps to exceed the mainly isolated view of datasets that dominates traditional DM.

PM provides new means of improving processes in a variety of domains. In our case, process-oriented knowledge discovery techniques in educational systems are an upcoming and emerging field of interest. PM applied to education data is called EPM. The combination of learning technology with PM brings considerable potential.[5] EPM involves the discovery, analysis and enhancement of processes and flows underlying the event logs generated by virtual learning environments (VLEs). However, we should bear in mind that there are limits to the value of an inductive, data-driven approach if there is no theory that can guide mining. Depending on the kind of process measures considered, students' navigation behavior could be relatively easily mined from a technical point of view. However, it should be used to assure initial efforts towards a guiding concept, rather than to start looking for regularities and patterns without a guiding conceptual framework. Taking this into consideration, EPM is able to build complete, compact educational process models that are able to reproduce all the observed behaviors, check to see if the modeling behavior matches the behavior observed, and project extracted information from the registrations in the pattern to make the tacit knowledge explicit, and to facilitate a better understanding of the process.[7] Likewise, the term Learnflow Mining in correspondence with Workflow Mining has been used by some authors[9,10] whereas many more authors[1,11–13] prefer the term EPM in correspondence with PM, which is the currently the most commonly used term.

There are also other related research areas used to discover learners' behavior (see Table 1). Next, we briefly address three of those which are most closely related to PM: intention mining (IM), sequential pattern mining (SPM), and graph mining (GM).

TABLE 1 | Main Related Areas with EPM

	Objectives	Algorithms	Models	Tools
Process mining	Discover underlying processes from event logs	Heuristic Miner, Fuzzy Miner, etc.	Petri Nets, Heuristic Net, BMMN, etc.	ProM, Disco, Celonis, etc.
Intention mining	Model the processes according to the purpose of the actors	Viterbi Algorithm, Baum-Welch Algorithm, etc.	KAOS, I*, Map, etc.	No tools found
Sequence pattern mining	Find common patterns between data examples where the values are delivered in a sequence	Generalized Sequential Patterns (GSP), Sequential Pattern Mining (SPAM), PrefixSpan, etc.	Sequences and subsequences, rules	SPFM, Himalaya Data Mining, etc.
Graph mining	Extract patters (sub-graphs) of interest from graphs that describe the underlying data	Branch-and-bound, On-line Plan Recognition, Recursive Matrix (R-MAT), etc.	Probabilistic graphs, signed graphs, colored graphs, Transition graphs, etc.	Graphviz, Deep Though, GSLAP, etc.

Intention Mining

IM or Intentional process mining is another recently emerged, related field of research. This field has the same objectives as PM but specifically addresses intentional process models, that is, processes focused on the reasoning behind activities.[14]

It is important to note that we have not found any research about the application of IM in education but the potential of this technique can be easily understood if it is particularly suitable to users' ways of thinking and working as it captures the human reasoning behind activities.

Sequence Pattern Mining

SPM[15] is a very commonly used technique in the DM environment for discovering common sub-sequences. Sequential pattern analysis aims to find relationships between occurrences of sequential events, that is, to find if any specific order of occurrences exists.[16] SPM is related to episode mining (EP); in fact, both techniques can be seen as variants of ASR. However, SPM methods find the most frequent event patterns across a set of event sequences, while EP discovers the most frequently used event patterns within a given sequence. There are other SPM-related techniques such as lag sequential analysis (LAS), t-pattern analysis and Markov models. While t-pattern analysis can be used to explore longer, more temporally separated sequences than LAS and Markov models, all these techniques are best suited to relatively short recurring sequences and analysis of event transitions.[17]

SPM techniques have been widely applied to analyze student learning behaviors. But, they are more indicated when trying to discover serial or simpler behavioral patterns than a process, for instance, learner activity paths in a course. So, SPM is not appropriate for discovering learning behaviors that describe the overall learning process.[18]

Graph Mining

GM is another popular pattern mining technique. The goal of GM or sub-graph mining is to find all frequent sub-graphs in a large graph or a database of graphs. GM and DM are closely related. Nevertheless, the main goal of graph mining is to supply new ideas and efficient algorithms for mining topological substructures embedded in graph data, while the main goal of DM is to supply ideas for mining and/or learning relational patterns represented by expressive logical languages. The former is more geometry-oriented and the latter more logic and relation oriented.[19] It is also important to differentiate between GM and social network analysis (SNA); SNA can be considered as an application of GM.

Graph-based Educational Data Mining (G-EDM) is a new, related, research area. Both G-EDM and EPM use graphs to represent information. However, while the task of GM is to extract patterns (sub-graphs of interest) from graphs that describe the underlying data and could be used further, for example, for classification or clustering, PM focusses on the process as a whole and therefore its graphs discover the overall learning process. Graphs are extremely important in the EDM community because of many types of data can be represented naturally as graphs including social network data and online discussions.

Finally, Table 1 shows a comparison of the previously described EPM related research areas.

FRAMEWORK AND CONCEPTS

An overview of the application of PM in the educational field is shown in Figure 1. In previous sections we have addressed how educational processes can be translated into executable processes by means of models. This EPM framework is an adaptation of the generic framework of PM[20] to the field of education[13,21] and below we describe the main components:

- **Educational world**: Basically two participants play an important role in any e-learning activity, teachers and students. Teachers supply appropriate resources to ensure student success. Students are the essential part of any e-learning activity, interacting with other participants (students or teachers), and with the system itself. Finally, courses, lectures, exams, etc. are used as resources for participants.
- **Virtual learning environment**: This supplies the basic structures and resources where the participants' learning actions and interactions occur. It also logs the events that occur during the e-learning process. Most provide teachers or researchers with basic tools for analyzing students' learning (marks evolution, number of activities done, forum participation, last log in, etc.) but do not provide specific tools that would allow educators to thoroughly assess the overall student learning process.
- **Event logs**: These are files that record events that occur in VLEs and are normally stored in databases. They contain a large amount of raw data about the educational agents' interaction with the VLE. Event logs need to be

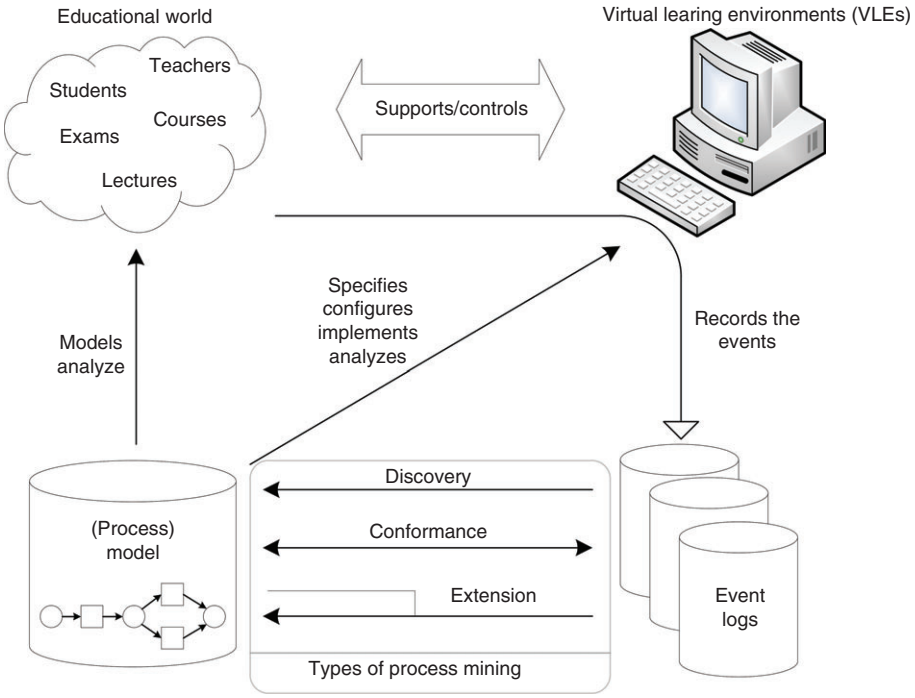


FIGURE 1 | EPM framework: types and components.

transformed into a particular file format in order to can be used by a specific PM tool.

- **Process models:** These uncover valuable information about how the participants of the educational world interact with the system starting from the event logs. They are obtained using different techniques in order to discover processes relevant to learning. Three main types of PM (see Table 2) can be distinguished[20]: discovery, conformance, and extension. These three basic types of PM can be also explained in terms of input and output (see Figure 2).

In addition to the three main types of PM, PM also provides distinct perspectives[22] such as control-flow, organizational, case, and time perspectives. The most commonly used in the educational environment is the control-flow perspective that focuses on the ordering of activities. The principal aim of this perspective is to discover an ideal description of all imaginable learning paths or trails in education[23] that can be generated when students navigate through a learning environment.

DATA AND TOOLS

In this section, we provide a deeper description of the data, potential difficulties, and software solutions used to perform EPM analysis.

TABLE 2 | Types of Process Mining

Type	Description	Application in Education
Process discovery	Constructs a comprehensive process model able to reproduce the behavior seen in the log file.	Instructor can visualize the behavior model of students' learning paths providing knowledge of the process instead of only the learning result
Conformance checking	Finds deviations between observed behaviors in event logs and generated process models.	Instructor can analyze whether the model corresponds to the behavior model in event logs and, for instance, find outliers.
Extension or enhancement	Aims to improve or extend a given process model based on information extracted from a specific event log related to the same process.	Instructor can detect bottlenecks or relationships between students in a course because different approaches can merged from a single integrated and extended process model.

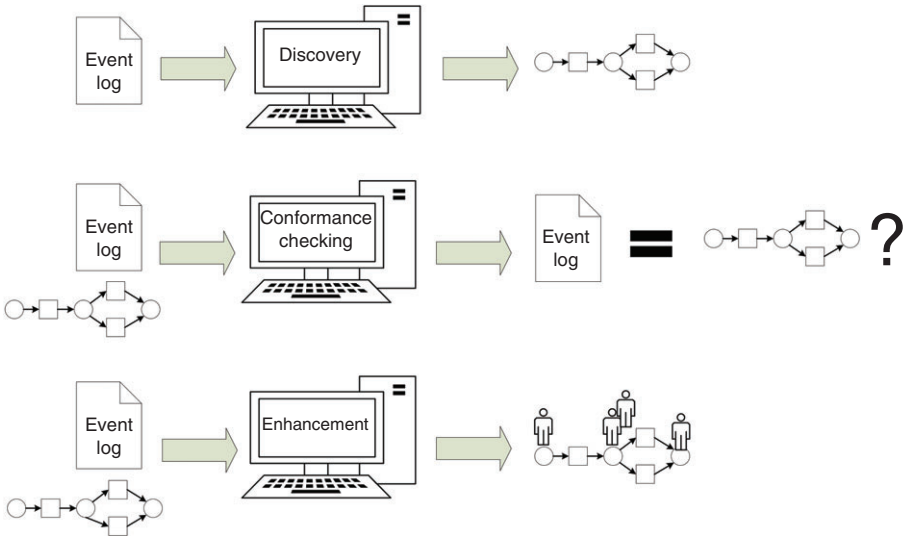


FIGURE 2 | Types of process mining explained in terms of input and output.

The starting point for PM is an event log.[22] An event log may be an Excel spreadsheet, a database table or a simple file that contains a trace/sequence of events. Each event is a row in the event log and refers to a case (case id), an activity (activity name), and a point in time (timestamp), sometimes it may contain additional information. Generally, they need to be transformed into specific formats for storing logs such as XES (eXtensible Event Stream) or MXML (Mining eXtensible markup language) in order to be used by a PM tool.[5] There are some specific tools supporting the conversion of different data sources to these formats such as ProMimport.[22]

Educational event logs can be gathered from a wide range of virtual e-learning environments such as Learning Management Systems (LMSs), Massive Open Online Courses (MOOCs), intelligent Tutoring Systems (ITSs), and Adaptive Hypermedia Systems (AHSs). See Figure 3 for an example of an event log generated from Moodle LMS. The Moodle system

Time	IP address	Full name	Action	Information
24/09/2013 12:22	150.214.10.12	Student53	Resource view	Tema 1
24/09/2013 12:24	150.214.10.12	Student53	Resource view	Tema 2
24/09/2013 12:25	150.214.10.12	Student53	Resource view	Tema 3
24/09/2013 12:26	150.214.10.12	Student53	Resource view	Tema 4
24/09/2013 12:28	150.214.10.12	Student53	Resource view	Clases prácticas (Diagnóstico)
24/09/2013 12:30	180.45.67.44	Student42	Resource view	Tema 1
24/09/2013 12:31	180.45.67.44	Student42	Resource view	Tema 2
24/09/2013 12:31	180.45.67.44	Student42	Resource view	Tema 3
24/09/2013 13:29	155.123.23.14	Student35	Folder view	Clases prácticas (Diagnóstico)
24/09/2013 13:29	155.123.23.14	Student35	Resource view	Tema 1
24/09/2013 13:29	155.123.23.14	Student35	Resource view	Tema 2
24/09/2013 13:29	155.123.23.14	Student35	Resource view	Tema 3
24/09/2013 13:29	155.123.23.14	Student35	Resource view	Tema 4
24/09/2013 14:06	145.124.25.65	Student49	Folder view	Clases prácticas (Diagnóstico)
24/09/2013 14:33	154.132.45.66	Student7	Folder view	Clases prácticas (Diagnóstico)
24/09/2013 14:33	154.132.45.66	Student7	Resource view	Tema 1
24/09/2013 14:33	154.132.45.66	Student7	Resource view	Tema 2
24/09/2013 14:33	154.132.45.66	Student7	Resource view	Tema 3
24/09/2013 14:33	154.132.45.66	Student7	Resource view	Tema 4

Timestamp
Case id
Activity name
Additional information

FIGURE 3 | Example of moodle event log.

logs every click (Time, IP, Student-FullName, Action and Additional-Information) that educational agents make for navigational purposes, generating a vast amount of *a priori* senseless information.

In general, several hurdles appear when handling event logs and they need to be overcome and borne in mind in EPM;[13,22] see Table 3 for further information.

Finally, many tools have emerged to support PM techniques such as:[22] ProM, Disco, Celonis Discovery, Perceptive Process Mining, QPR ProcessAnalyzer, Aris Business Process Analysis, Fujitsu Process Analytics, XMAnalyzer, and StereoLOGIC Discovery Analyst. However, all of them are general purpose PM tools and only a few have been used in EPM (see Table 4).

Only three of these PM tools have been referenced in all of the papers (see Table 4). ProM tool is a generic open-source framework for implementing PM and it is the most complete and most commonly used in EPM, followed by Disco, which is also a general purpose tool but commercial. There is only one PM software specific to the education domain, named SoftLearn.[24] It provides a graphical interface that teachers can use to visualize learning paths as activity graphs and to access the relevant data generated in the learning activities.

TECHNIQUES

In this section, we describe the most commonly used techniques in EPM. We highlight four main groups

TABLE 3 | Challenges when handling event logs

Issue	Description	Example in EPM
Correlation	Events are grouped per case in an event log. Events need to be related to each other.	Students perform similar kinds of actions in a forum.
Noise	An event log may contain outliers. Exceptional behavior is not representative of typical behavior of the process.	Students can leave an open session.
Incompleteness	A common problem is that the event log contains too few events to be able to discover some of the underlying control-flow structures.	E-learning systems fall down.
Distribution	Data may be distributed over a variety of sources.	Student information can be gathered from different sources: Administrative information, theory and practice classroom, online learning environments, etc.
Timestamp	Events need to be ordered per case.	Typical problems: only dates, different time zones, delayed logging
Snapshots	Cases may have a lifetime extending beyond the recorded period	Student case was started before the beginning of the event log.
Scoping	What is the *process* we want to investigate? How to decide which tables to include?	LMS and MOOC can provide different tables to investigate different process.
Granularity	The events in the event log are at a different level of granularity.	Educational information may have different levels of granularity, ranging from low level clicks, to activities, courses, etc.
Contextualization	Events occur in a particular context. This context may explain certain phenomena. This requires the merging of event data with contextual data	Teachers discover models in a repeat student class.
Size	Number of cases or events in event logs can be high. These files can be difficult to handle due to their size.	Virtual learning environments can generate huge logs.
Complexity	Distinct traces and activities in event logs can be high due to the large diversity of behaviors in students' learning paths.	Educational environments can generate complex models that are difficult to understand, named spaghetti models.
Concept drift	Situation in which the process changes while being analyzed.	Courses and study curriculums may be modified at any time during learning span.
Privacy	Privacy and authentication has many ethical dimensions.	Students need to be aware what the system is doing with their data.

TABLE 4 | Comparison between the Main Tools Used in EPM

Company (Country)	ProM Eindhoven Technical University (Netherlands)	Disco Fluxicon (Netherlands)	SoftLearn University of Santiago de Compostela (Spain)
Purpose	General	General	Specific (education)
Type	Free	Commercial	Private
Filtering	Yes	Yes	No
Process discovery	Yes	Yes	Yes
Conformance checking	Yes	No	No
Social network mining	Yes	No	No
Number of papers/Works	21	7	1

of techniques: discovery, conformance checking, dotted chart analysis, and SNA.

Discovery Techniques

Process discovery techniques build a process model based solely on an event log by capturing the behavior seen in the log. They focus on the control-flow perspective of process. There are a lot of algorithms in PM for discovering underlying processes from event logs, but the most often used in educational domains are (see Table 5):

- **Alpha algorithm:** a relatively intuitive and simple technique based on dependency relation between events which requires ideal event logs without noise. It was one of the first algorithms that was able to deal with concurrency.[25]
- **Heuristic Miner algorithm:** this uses likelihood by calculating the frequencies of relations between the tasks (e.g., causal dependency, loops, etc.) and constructs dependency/frequency tables and dependency/frequency graphs.[14] The Heuristic Miner algorithm was designed to make use of a frequency based metric and so is less sensitive to noise and the incompleteness of logs.[26]
- **Genetic algorithm:** this provides process models built on causal matrixes (input and output dependencies for each activity). This approach tackles problems such as noise, incomplete data, non-free-choice constructs, hidden activities, concurrency, and duplicate activities.[14]
- **Fuzzy miner:** this is one of the newer process discovery algorithms. It is the first algorithm to directly address the problems of large numbers of activities and highly unstructured behavior.[27]

TABLE 5 | Representation Models Used in EPM Works

Work/Paper	Petri Nets	HLPN	Fuzzy	Heuristic
Weijters et al.[4]	X			X
Günther and Van Der Aalst[27]			X	
Pechenizkiy et al.[20]	X		X	X
Reimann et al.[17]	X			X
Trcka and Pechenizkiy[7]		X		
Southavilay et al.[39]				X
Trcka et al.[1]	X		X	
Poncin et al.[49]			X	
Schoor and Bannert[37]			X	
Anuwatvisit et al.[48]	X			
Ayutaya et al.[47]	X			X
Bergenthum et al.[9]	X	X		
van der Aalst et al.[12]			X	
Reimann et al.[8]			X	
Bannert et al.[18]	X		X	
Cairns et al.[40]				X
Cairns et al.[32]			X	
Bogarin et al.[26]				X
Cairns et al.[31]	X			X
Cairns et al.[13]			X	X
Mukala et al.[34]			X	
Ariouat et al., 2016[43]				X
Doleck et al.[41]			X	
Okoye et al.[55]			X	
Sedrakyan et al.[54]			X	
Vahdat et al.[42]			X	
Vidal et al.[21]	X			

Good notation is necessary in order to represent ready process models to the end-user. All the above mentioned algorithms produce a process model that is normally independent of the desired representation. There are different types of representations or metamodels in PM such as Petri nets, Workflow nets, Fuzzy nets, Heuristic nets, Causal nets, Process tree, BPMN (Business Process Model and Notation), EPC (Event Driven Process Chain), and UML (Unified Modeling Language) Activity Diagram. Although Petri nets and BPMN are the most often used in PM,[14] the most commonly used in education domain are (see Table 5):

- **Petri Nets:** Graphs with two types of nodes linked by directed arcs. The first type of node is known as place and is represented by an ellipse. Places can store a multi-set of values, called tokens. Transitions are represented as rectangles and identify active elements of the net.[28]
- **High-level Petri Net (HLPN):** Extended classical Petri Nets with color, time and hierarchy. Colored Petri Nets (CPN) were the first concrete realization of HLPN and were a graphical language for analyzing the properties of concurrent systems.[14]
- **Fuzzy net:** Simplifies the complete model by preserving highly significant events or edges, aggregating less significant but highly correlated edges and nodes by clustering, and abstracting from less significant and poorly correlated edges and nodes by removing them from the simplified model.[27]
- **Heuristic nets:** A directed cycle graph which represents the most frequent behaviors of the students in the dataset used. In heuristic nets the square boxes represent the actions and the arcs/links represent dependences/relations between actions.[26]

Additionally, it is possible to automatically transform a model from one representation/notation to another when using some PM software. In Figure 4, we show two different representations/notations obtained from the same log file. A Petri Net showing the causality and parallelism of the events and a Heuristic Net showing the frequency of the events and how strong the dependency between events is.

It is necessary to state that the most commonly used representation model in EPM research is the Fuzzy net, followed by Petri net and Heuristic net, with HLPN the least used (see Table 5).

Conformance Checking Techniques

The goal of conformance checking is to find commonalities and discrepancies between the modeled behavior and the observed behavior. In the EPM literature, two techniques stand out in conformance checking:

- Linear Temporal Logic (LTL) Checker. This checks whether the logs satisfy some Linear Temporal Logic (LTL) formula.[29] LTL Checker does not compare a model with the log, but a

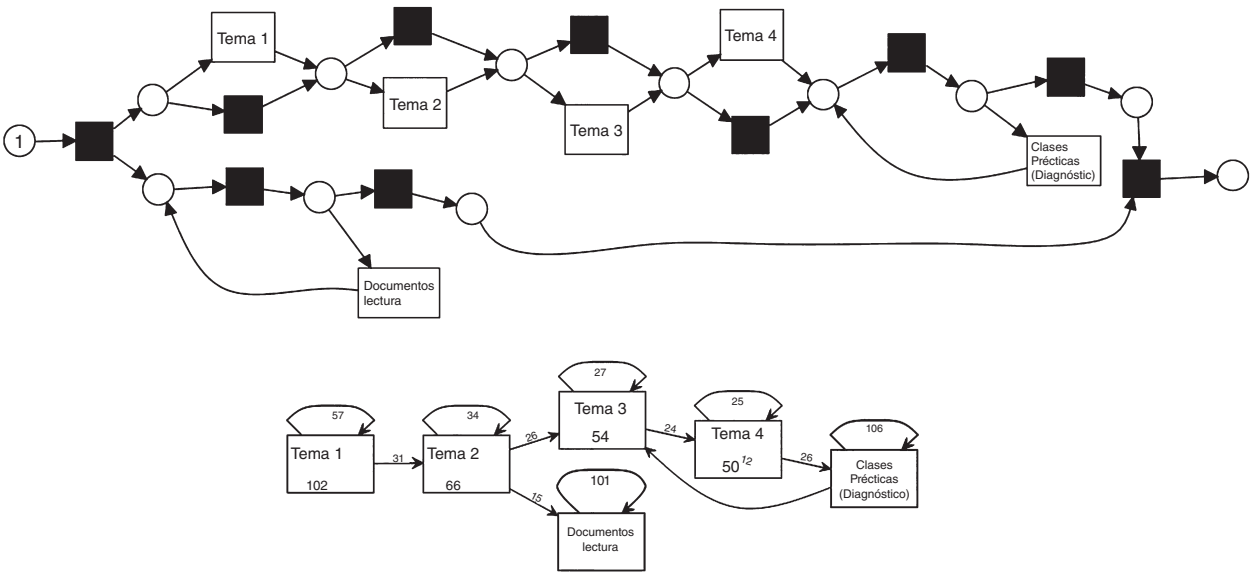


FIGURE 4 | Examples of Petri and Heuristic Net generated with the same log data.

set of requirements described by the (linear) temporal logic LTL.

- The conformance checker. This requires a model in addition to an event log. It replays an event log within a Petri net model in a non-blocking way while gathering diagnostic information that can be accessed afterwards.[30]

Dotted Chart Analysis Technique

A dotted chart shows the spread of events over time by plotting a dot for each event in an event log thus providing some insight into the underlying process, its performance and any patterns of interest. It represents the log file visually, showing a general time perspective of the process. The chart has two orthogonal dimensions: time and component types. The time is measured along the horizontal axis of the chart. The component types are shown along the vertical axis.[31] Figure 5 shows an example of dotted chart about the daily work carried out by students in Moodle. Each row is a different task or Moodle event in the course, and the size of dots represent how many students have done this task at a particular time.

Social Network Analysis Technique

SNA refers to the collection of methods, techniques and tools in sociometry aimed at the analysis of social networks. SNA aims to extract social networks from event logs based on the observed interactions between performers, depending on how process instances are routed between these performers.[32] A social network consists of nodes representing organizational entities and arcs representing relationships. Figure 6 shows an example of social networks that represents how and how much students interact in a Moodle forum. Bigger nodes represent more active students and Arcs represent the moment when they interact.

Finally, a summary of the most commonly used techniques in EPM research.

As Table 6 shows, the most commonly used discovery algorithms are Heuristic Miner and Fuzzy Miner. Conformance checker is the most commonly used conformance technique. The Dotted Chart is more commonly used in educational research than SNA.

APPLICATION DOMAINS

EPM has been used in a wide range of application domains in education in order to address varying educational problems. In this section, we describe the main body of literature, giving special importance to current EPM applications rather than the specific results.

MOOCs, LMS and Hypermedia Environments

Massive Open Online Courses (MOOCs), Learning Management System (LMS), Hypermedia and other similar online learning environments supply free learning opportunities to the online community. Log files generated by these systems provide an insight into how people follow the course, when they watch videos or lectures, and when they hand in tasks, among others.

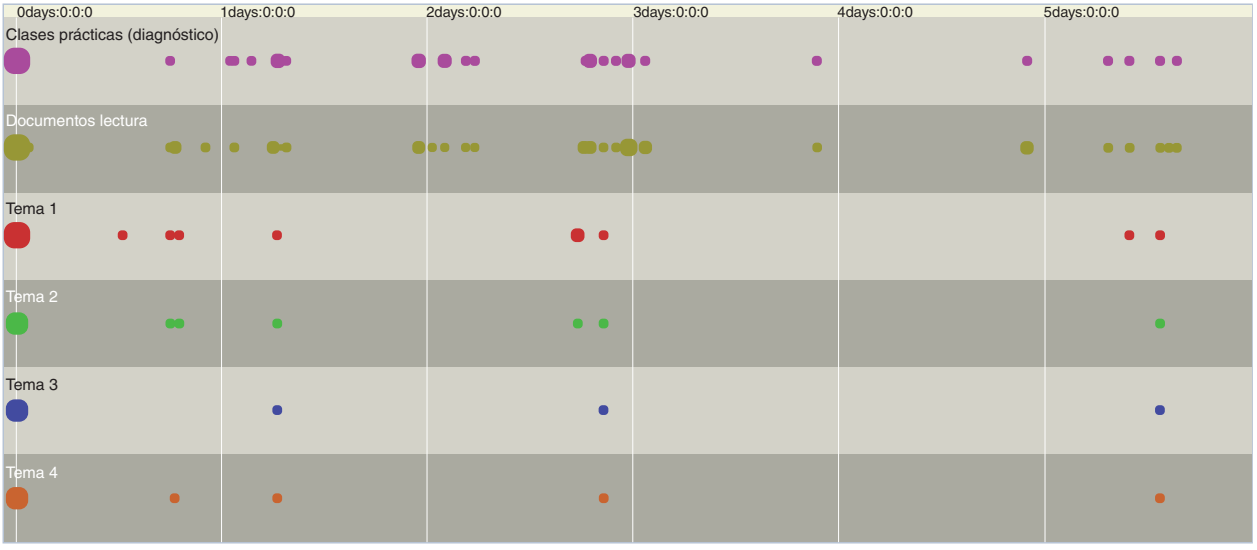


FIGURE 5 | Example of a dotted chart of the daily work carried out by students in Moodle.

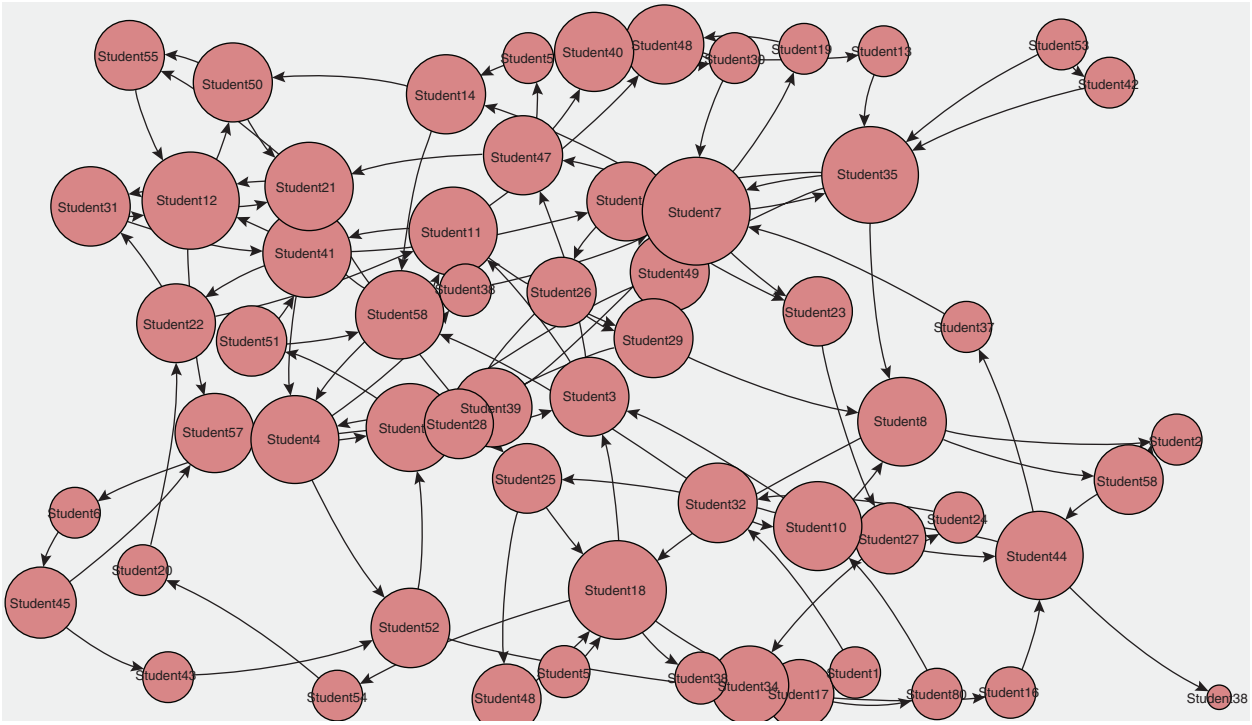


FIGURE 6 | Example of a social network that represents how and how much students interact in a Moodle forum.

There is a lot of research about the application of PM in this type of learning environment. Trcka et al.[1] exemplified the applicability of PM and discussed some of its potential for extracting knowledge from LMSs, considering only students´ examination traces. In Bogarin et al.,[26] the authors used data from Moodle logs and proposed using clustering in order to be able to obtain more specific and accurate Process Models of students' behavior. In a similar environment Reiman et al.[8] proposed the use of traces to study self-regulated learning (SRL) in a hypermedia environment based on theoretical principles and PM. Using those principles, Bannert et al.[18] detected differences in frequencies of SRL events using PM techniques and found that successful students demonstrate more learning and regulation events. In other research Mukala et al.[33] used PM techniques in order to trace and analyze students' learning habits based on MOOC data. Results indicated that successful students follow a sequentially-structured watching pattern while unsuccessful students are unpredictable and have poorly structured processes. In later research Mukala et al.[34] made use of alignment-based conformance checking to extract and analyze students' learning patterns in a MOOC. Along similar lines, Emond and Buffett[35] applied process discovery mining and sequence classification mining techniques to model and support SRL in heterogeneous environments of learning content, activities, and social networks. They used a data set of semi-structured learning activities taken from the DataShop at the Pittsburgh Science of Learning Centre. Finally, Vidal et al.[21] used logs from a VLE to extract the learning flow structure using PM, and to obtain the underlying rules that control students' adaptive learning by means of decision tree learning.

Computer-Supported Collaborative Learning

Computer-supported collaborative learning (CSCL) is characterized by the sharing and construction of knowledge between participants using technology as their primary means of communication or as a common resource.

PM has been applied in CSCL in order to supply feedback to students on their decision-making processes.[17] The goal was to use PM to identify models of decision-making groups that took place in a chat room. In a similar study, Bergenthum et al.[9] proposed a modeling language for collaborative learnflows that specifically takes the actors, the roles, and the explicit representation of groups into account. Their research extends on previous work focused on the discovery of the structures for flow control using methods from the area of WM, with

TABLE 6 | Techniques used in EPM research

Work/Paper	Discovery Algorithm	Conformance Techniques	Dotted Chart	SNA
Weijters et al.[4]	Heuristic Miner			
Pechenizkiy et al.[20]	Heuristic Miner, Fuzzy Miner	Conformance checker	X	
Reimann et al.[17]	Heuristic Miner			
Trcka and Pechenizkiy[7]		Conformance checker		
Southavilay et al.[39]	Heuristic Miner		X	
Trcka et al.[1]	Fuzzy Miner	LTL—Conformance checker	X	
Poncin et al.[49]	Fuzzy Miner		X	
Ayutaya et al.[47]	Heuristic Miner			
Anuwatvisit et al.[48]		Conformance checker		
Schoor and Bannert[37]	Fuzzy Miner			
van der Aalst et al.[12]	Fuzzy Miner	Conformance checker	X	
Reimann et al.[8]	Fuzzy Miner			
Barreiros et al.[24]	Genetic Algorithm			
Bannert et al.[18]	Fuzzy Miner	LTL—Conformance checker		
Cairns et al.[40]	Heuristic Miner	LTL		
Cairns et al.[32]	Fuzzy Miner			X
Bogarin et al.[26]	Heuristic Miner			
Cairns et al.[31]		LTL—Conformance checker	X	
Cairns et al.[13]	Fuzzy Miner	LTL—Conformance checker	X	X
Mukala et al.[34]	Fuzzy Miner	Conformance checker	X	
Vahdat et al.[42]	Fuzzy Miner			
Ariouat et al.[43]	Heuristic Miner			
Okoye et al.[55]	Fuzzy Miner			
Sedrakyan et al.[54]	Fuzzy Miner		X	
Vidal et al.[21]	Genetic Algorithm			

support for dynamic role assignment.[36] Other authors like Schoor and Bannert[37] have explored sequences of social regulatory processes during a CSCL task and their relationship to group performance. This study used PM to identify process patterns for high and low group-performance dyads. In the most recent research in this domain, Porouhan and Premchaiswadi[38] apply several PM techniques such as social network mining, basic performance analysis, role hierarchy mining, and dotted chart analysis with the aim of increasing the instructor's awareness/knowledge about the collaborative dynamics in each group. A particular application of EPM to this domain is Collaborative Writing (CW). CW is widely used in education environments where students often use computers to take notes during lectures and write essays for their assignments. Thanks to the availability of the Internet, students can also write collaboratively by sharing their documents in a number of ways. PM has been used in Southavilay et al.[39] to analyze students' writing processes and how these processes correlate to the quality and semantic features of the final product. They used documents collected from different groups of undergraduate students writing collaboratively in order to evaluate the proposed heuristics and illustrate the applicability of PM techniques to analyzing the writing process.

Professional Training

Education and training centers have made their professional training courses more agile in order to respond to the changing needs of the job market and meet time-to-skill requirements.[32]

PM has been used in different types of professional training courses. Cairns et al.[32] showed how PM can be used to monitor and improve educational processes in the field of professional training. Their research aims to develop generic methods which could be applied to general education issues and more specific applications concerning professional

TABLE 7 | Targets in EPM Educational Application Domains

Application	Work/Paper	Target
MOOCs, LMS, and hypermedia environments	Mukala et al.[34]	To detect learning difficulties
	Mukala et al.[33]	To generate recommendations or advice for students.
	Bogarin et al.[26]	To gain a better understanding of the underlying educational process
	Vidal et al.[21]	To improve management of learning objects
	Bannert et al.[18]	To detect learning difficulties and discover sequential patterns
	Reimann et al.[8]	To discover sequential patterns
	Trcka et al.[1]	To discover learning flows
	Emond and Buffett.[35]	
Computer-supported collaborative learning	Reimann et al.[17]	To discover learning flows and provide feedback
	Bergenthum et al.[9]	To discover learning flows
	Schoor and Bannert[37]	To discover sequential patterns
	Porouhan and Premchaiswadi[38]	To gain a better understanding of the underlying educational process
	Southavilay et al.[39]	To gain a better understanding of the underlying educational process and detect learning difficulties
Professional training	Cairns et al.[32]	To analyze social networks
	Cairns et al.[31]	To discover learning flows
	Doleck et al. 2016[41]	To gain a better understanding of the underlying educational process and detect learning difficulties
	Vahdat et al.[42]	
	Ariouat et al., 2016[43]	To gain a better understanding of the underlying educational process
Curriculum mining	Trcka and Pechenizkiy[7]	To gain a better understanding of the underlying educational process
	Wang and Zaïane[44]	To gain a better understanding of the underlying educational process and generate recommendations or advice for students.
	Schulte et al.[45]	To generate recommendations or advice for students.
Computer-based assessment	Pechenizkiy et al.[20]	To provide feedback
	Tóth et al.[46]	To detect learning difficulties
Student registration	Anuwatvisit et al.[48]	To gain a better understanding of the underlying educational process
	Ayutaya et al.[47]	To discover learning flows
Software repositories	Poncin et al.[50]	To gain a better understanding of the software development process
	Poncin et al.[49]	
Structured inquiry cycle	Howard et al.[51]	To detect learning difficulties
	Jeong et al.[52]	
3D educational virtual worlds	Fernández-Gallego et al.[53]	To discover learning flows

training or e-learning fields for the extraction, analysis, enhancement and personalization of educational processes. In similar research, Cairns et al.[31] analyzed training processes and their conformance with established curriculum constraints, educators' hypotheses and prerequisites, and to enhance training process models with performance indicators such as execution time, bottlenecks and decision points. Doleck et al.[41] explored knowledge-based discovery approaches to understand learner behaviors in a medical computer-based learning environment using PM techniques in order to provide a more coherent

picture about clinical diagnosis reasoning. Vahdat et al.[42] carried out their study in a simulation environment for learning digital electronics. They exploited PM methods to investigate and compare the learning processes of professionals. In order to do that, they measured the understandability of their process models through a complexity metric. Additionally, Ariouat et al.[43] tried to identify the best training paths in real-world professional training databases from a global consulting company.

Curriculum Mining

A curriculum is partially designed by an educational institution in order to accomplish certain goals. Curricula normally suggest that students can follow differing paths from start to end due to a liberal approach in selecting courses.[44]

A domain-driven EPM approach was proposed by Trcka and Pechenizkiy[7] in curriculum mining. They proposed a framework which assumes that a set of pattern templates can be predefined to focus the mining in a desired way and make it more effective and efficient. The framework is aimed at helping educators analyze educational processes based on formal modeling. In other related research, Wang and Zaïane[44] discovered a curriculum process model of students taking courses and compared the paths that successful and less successful students tended to take, highlighting discrepancies between them. In other work Schulte et al.[45] presented research into EPM and student data analytics in a whole university scale approach with the aim of providing insight into questions raised by degree pathways. Their goal was to uncover statistically significant and meaningful patterns in students' course pathway choices, and to provide student support units, degree and course coordinators with longitudinal indicators that could be used to inform students.

Computer-based Assessment

Computer-based assessment (CBA) is, in essence, the practice of giving quizzes and tests on the computer instead of using traditional pencil and paper formats. Computer based assessment is widely used in many different VLEs.

PM has been used to exclusively analyze assessment data coming from two online multiple choice test studies showing the utility of process discovery, conformance checking and performance analysis techniques.[20] In a similar context, Tóth et al.[46] described how to extract process-related information from event logs, and how to use these data in problem-solving assessments and describe methods which help discover novel information based on individual problem-solving behavior.

Student Registration

Student registration deals with all the requirements and steps of the academic or course registration process. It is critical to check on automated management system processes in the educational domain in order to produce expected results in terms of quality and timely student registration processes.[47]

In this context, Ayutaya et al.[47] used the Heuristic Mining technique to gain insight into student registration processes at a Thai university. The most important characteristic of Heuristics Miner is its robustness against noise and exceptions. Because Heuristics Miner is based on the frequency of patterns it is possible to focus on the main behavior in the event log and that makes it especially appropriate for unstructured educational processes. Anuwatvisit et al.[48] used Conformance checker in order to detect discrepancies between the flows prescribed in a student registration model and the actual process instances. In addition, they extended the models with performance characteristics and business rules.

Software Repositories

Developers and development teams are involved in software development processes, often from different locations. In these projects, different kinds of software repositories such as source-code management systems, document repositories, mail archives, bug trackers and version control systems are used to support communication and coordination.

PM has also been applied to mining software repositories. Poncin et al.[49] identified the challenges that need to be addressed to enable this application, discussing how they can be addressed and presented through FRASR (Framework for Analyzing Software Repositories). PM has been also applied in Poncin et al.[50] to analyze data from multiple software repositories. The preprocessing step extracts information from software repositories, which have different structures, and combines information into an event log, while the analysis step is aimed at discovering the process structure, reflected in the log, and analyzing whether it is correct or visualizing it.

Structured Inquiry Cycle

A structured inquiry cycle is a kind of adaptation strategy that combines explicit structuring and scaffolding with increased learner control to create a freer, more personalized learning experience due to high potential

variation in prior knowledge, metacognitive skills, and motivation within learner populations.[51]

A PM approach using a structured inquiry cycle has been applied in a corpus of online modules for adult informal learners.[51] Informal learning situations often exhibit high variability within the learner population, especially when learning experiences and environments offer broad availability. However, this freedom of navigation sometimes has negative effects on learning experiences, particularly when prior domain knowledge or learning skills are weak. The authors demonstrated that Petri Net process models which aid collaborative planning and reviewing results, where unshared, informal understanding can be a hindrance.

In a similar context, Jeong et al.[52] used a hidden Markov model approach for exploratory sequence analysis by applying the methodology to studying student learning behaviors in a new domain that promotes an inquiry cycle.

3D Educational Virtual Worlds

3D Educational Virtual Worlds are environments that encourage interaction between students and teachers. These environments encourage students (as avatars) to perform learning activities that were not initially scheduled by the teachers.

PM has been used in order to find out what is happening in 3D student learning processes. With that aim, Fernández-Gallego et al.[53] presented a learning analytics framework for 3D educational virtual worlds that focuses on discovering learning flows and checking conformance through PM techniques because, in this specific domain, the interactions among students are continuous and there is a lot of noise, in other words, a high number of activities that are not significant from a pedagogical point of view.

Finally, Table 7 shows a summary of all the previously described EPM research and its target grouped by application domain. On the one hand, we can see that currently, the most active domains are: Research into MOOC, LMS and Hypermedia environments, discovering in CSCL, and discovering in Professional Training. On the other hand, the results of EPM can be used to gain a better understanding of the underlying educational processes, to provide feedback to students, teachers and researchers, to detect learning difficulties and to help students with specific learning disabilities, to improve management of learning objects, to generate advice for students, among many other uses. Table 7 shows the general target of every study indicating that the most frequent targets in current EPM research are focused on gaining a better understanding of underlying educational processes, detecting learning difficulties and discovering learning flows.

CONCLUSION AND FUTURE WORKS

This paper presents a comprehensive introduction to EPM, one of the most promising EDM techniques. EPM is young, emerging field which is closely related to other research areas such as IM, SPM and G-EDM. The present work describes the EPM framework which is an adaptation of the generic PM framework and addresses the different hurdles that may arise when handling event logs, the most commonly used techniques, and most often used tools. Additionally, it provides an overview of the main research to date that aims to be a guide to the available research in this regard.

EPM allows a better understanding of the underlying educational process from raw event data but as an emerging area, it faces many challenges and has many prospects for the future. We would like to outline some important issues that will be especially challenging in EPM's near future, such as:

- **Development of more specific EPM tools** in order to bring PM closer to the domain experts (i.e., educational specialists and researchers, who do not necessarily have all the technical background), helping them to analyze educational processes. SoftLearn[24] is currently the only specific EPM tool. Additionally, PHIDIAS is a project developed by ALTRAN researchers[32] with the aim of developing an interactive platform tailored for educational process reconstruction and analysis. This platform will allow different education centers and institutions to load their data and provide access to advanced DM and PM services.
- **Using semantics to improve EPM.** Semantic concepts can be layered on top of existing learner information assets to provide a more conceptual analysis of real time processes capable of providing real world answers that are closer to human understanding. Along these lines, Cairns et al.[32] proposed how linking labels in event logs to their underlying semantics can bring educational process discovery to the conceptual level. In this way, more accurate and compact educational processes can be mined and analyzed at different levels of abstraction. Sedrakyan et al.[54] focused on the modeling activities that can potentially affect the semantic quality of a conceptual model process. They constructed a semantically correct conceptual model that reflects the structural and

dynamic view of a given domain description. Finally, Okoye et al.[55] proposed semantic PM to enrich streams of event data logs from a learning process using semantic descriptions that references concepts in an ontology that is specifically designed for representing learning processes.

- **Using recommendation in EPM.** The information discovered in PMs has to be, not only comprehensible, but also useful for the end-users' decision-making. For example, instead of showing the overall process model obtained, it is better to abstract its representation using, for example, the normal academic notation understandable to end-users[8,18] or in the form of a list of suggestions, recommendations, and conclusions about the results.[13] So, in addition of the three main types of EPM, some authors,[14,32] consider recommendation as another emerging type. Recommendation can be seen an extension of the enhancement type instead of a new type of EPM. Chen et al.[56] used PM to extract and reorganize the learning process in order to be able to recommend a learning unit for the target user to learn as the next step. Wang and Zaïane,[44] used PM to adjust the requirements for the curriculum and to recommend courses to students based on expected outcome. Cairns et al.[31] show recommendations of the best course units or learning paths to students (depending on their profiles, preferences or target skills) and the on-line detection of unmet prerequisites.
- **The application of EPM to other emergent educational domains** such as games, mobile, and ubiquitous learning environments. The play traces resulting from the learner's activity in learning games are hard for teachers to analyze and interpret. EPM techniques have started to be applied in Learning Games applications[57] in order to analyze of student behaviors. Digital mobile devices such as tablets, PDAs, and smart phones are also being used increasingly often for educational purposes. Process Mining and Learner Behavior Analytics[38] have been used in a Collaborative and Web-Based MultiTabletop Environment.
- **Make more free EPM datasets public** in order to test theoretical models and assumptions by methods of PM. Free EPM datasets could be very useful for testing some *ad-hoc* models. However, not all the current EPM datasets used in research are available to download. We believe it is essential that EPM datasets are freely available in order for EPM research to grow much more quickly. In fact, we have only found one free specific EPM dataset (https://archive.ics.uci.edu/ml/datasets/Educational+Process+Mining+(EPM)%3A+A+Learning+Analytics+Data+Set). Therefore, one of the most important next steps is to promote the sharing of online data-collections that can be analyzed from multiple perspectives, with various methods and tools.

ACKNOWLEDGMENTS

This research is supported by project TIN2014-55252-P from the Spanish Ministry of Science and Technology and the Department of Science and Innovation under the National Program for Research, Development, and Innovation: EDU2014-57571-P.We have also received funds from the European Union, through European Regional Development Funds (ERDF); and the Principality of Asturias, through its Science, Technology and Innovation Plan (grant GRUPIN14-053).

REFERENCES

1. Trcka N, Pechenizkiy M, Van der Aalst WMP. *Process Mining from Educational Data* (Chapter 9); 2011.
2. Van der Aalst W, Weijters T, Maruster L. Workflow mining: discovering process models from event logs. *IEEE Trans Knowl Data Eng* 2004, 16:1128–1142.
3. Romero C, Ventura S. Educational data science in massive open online courses. *WIREs Data Mining Knowl Discov* 2017, 7:1–12.
4. Weijters AJMM, van Der Aalst WM, De Medeiros AA. Process mining with the heuristics miner-algorithm. In: *Technische Universiteit Eindhoven, Tech. Rep. WP*. Vol 166, 2006, 1–34.
5. Romero C, Cerezo R, Bogarín A, Sánchez-Santillán M. Educational process mining: a tutorial and case study using moodle data sets. In: *Data Mining and Learning Analytics: Applications in Educational Research*. Hoboken, NJ: John Wiley & Sons; 2016, 1–28.
6. Mans RS, van der Aalst W, Vanwersch RJ. *Process Mining* in Healthcare: Evaluating and Exploiting Operational Healthcare Processes. Berlin, Germany: Springer; 2015, 17–26.

7. Trcka N, Pechenizkiy M. From local patterns to global models: towards domain driven educational process mining. In: *Ninth International Conference on Intelligent Systems Design and Applications*, IEEE, Pisa, Italy, 2009, 1114–1119.
8. Reimann P, Markauskaite L, Bannert M. E-research and learning theory: what do sequence and process mining methods contribute? *Br J Educ Technol* 2014, 45:528–540.
9. Bergenthum R, Desel J, Harrer A, Mauser S. Modeling and mining of learnflows. *Trans Petri Nets Other Models Concurrency* 2012, 5:22–50.
10. Perez-Rodriguez R, Caeiro-Rodriguez M, Anido-Rifon L. Enabling process-based collaboration in Moodle by using aspectual services. In: *Ninth IEEE International Conference on Advanced Learning Technologies*, IEEE, Riga, Latvia, 2009, 301–302.
11. Romero C, Ventura S. Data mining in education. *WIREs Data Mining Knowl Discov* 2013, 3:12–27.
12. Van der Aalst WM, Guo S, Gorissen P. Comparative process mining in education: An approach based on process cubes. In: *International Symposium on Data-Driven Process Discovery and Analysis*. Springer Berlin Heidelber, Riva del Garda, Italy; 2013, 110–134.
13. Cairns AH, Gueni B, Assu J, Joubert C, Khelifa N. Process mining in the education domain. *Int J Adv Intell Syst* 2015, 8:219–232.
14. Khodabandelou G, Hug C, Deneckere R, Salinesi C. Process mining versus intention mining. In: *Enterprise, Business-Process and Information Systems Modeling*. Springer Berlin Heidelberg,Valencia, Spain; 2013, 466–480.
15. Agrawal R, Srikant R. Mining sequential patterns. In: *Data Engineering Proceedings of the Eleventh International Conference*. IEEE, Taipei, Taiwan; 1995, 3–14.
16. Nesbit JC, Zhou M, Xu Y, Winne P. Advancing log analysis of student interactions with cognitive tools. In: *12th Biennial Conference of the European Association for Research on Learning and Insruction (EARLI)*. Budapest, Hungary; 2007, 2–20.
17. Reimann P, Frerejean J, Thompson K. Using process mining to identify models of group decision making in chat data. In: *Proceedings of the 9th international conference on Computer supported collaborative learning, Vol. 1*. Rhodes, Greece; 2009, 98–107.
18. Bannert M, Reimann P, Sonnenberg C. Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacogn Learn* 2014, 9:161–185.
19. Washio T, Motoda H. State of the art of graph-based data mining. *Acm Sigkdd Explor Newslett* 2003, 5:59–68.
20. Pechenizkiy M, Trcka N, Vasilyeva E, van Aalst W, De Bra P. Process mining online assessment data. In: *Educational Data Mining*. Córdoba, Spain; 2009, 279–288.
21. Vidal JC, Vázquez-Barreiros B, Lama M, Mucientes M. Recompiling learning processes from event logs. *Knowledge-Based Syst* 2016, 100:160–174.
22. Van der Aalst WM. *Process Mining: Data Science in Action*. Berlin, Germany:Springer; 2016.
23. Schoonenboom J, Levene M, Heller J, Keenoy K, Turcansyi M. *Trails in Education: Technologies that Support Navigational Learning*. Rotterdam, The Netherlands: Sense Publishers; 2007.
24. Barreiros BV, Lama M, Mucientes M, Vidal JC. Softlearn: a process mining platform for the discovery of learning paths. In: *14th International Conference on Advanced Learning Technologies*. IEEE, Athens, Greece; 2014, 373–375.
25. Mekhala A. Review paper on process mining. *Int J Eng Res Technol* 2015, 1:11–17.
26. Bogarín A, Romero C, Cerezo R, Sánchez-Santillán M. Clustering for improving educational process mining. In: *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. ACM, Indianapolis, USA; 2014, 11–15.
27. Günther CW, Van Der Aalst WM. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In: *International Conference on Business Process Management*. Springer Berlin Heidelberg; Brisbane, Australia; 2007, 328–343.
28. Vidal JC, Lama M, Bugarín A. Petri net-based engine for adaptive learning. *Expert Syst Appl* 2012, 39:12799–12813.
29. Van Dongen BF, de Medeiros AKA, Verbeek HMW, Weijters AJMM, Van Der Aalst WM. The ProM framework: a new era in process mining tool support. In: *International Conference on Application and Theory of Petri Nets*. Springer Berlin Heidelberg; Miami, USA; 2005, 444–454.
30. Rozinat A, van der Aalst WM. Conformance testing: Measuring the fit and appropriateness of event logs and process models. In: *International Conference on Business Process Management*. Springer Berlin Heidelberg; Nancy, France; 2005, 163–176.
31. Cairns AH, Gueni B, Assu J, Joubert C, Khelifa N. Analyzing and improving educational process models using process mining techniques. In: *The Fifth International Conference on Advances in Information Mining and Management*. Brussels, Belgium; 2015, 17–22.
32. Cairns AH, Gueni B, Fhima M, Cairns A, David S, Khelifa N. Towards custom-designed professional training contents and curriculums through educational process mining. In: *The Fourth International Conference on Advances in Information Mining and Management*. Paris, France; 2014, 53–58.
33. Mukala P, Buijs J, van der Aalst WMP. Uncovering learning patterns in a MOOC through conformance alignments. In: Tech. rep., Eindhoven University of Technology, BPM Center Report BPM; 2015.

34. Mukala P, Buijs J, Leemans M, van der Aalst W. Learning analytics on coursera event data: a process mining approach. In: *Proceedings of the 5th International Symposium on Data-driven Process Discovery and Analysis*. Vienna, Austria; 2015, 18–32.
35. Emond B, Buffett S. Analyzing student inquiry data using process discovery and sequence classification. In: *Proceedings of the 8th International Conference on Educational Data Mining*. Madrid, Spain; 2015, 412–415.
36. Bergenthum R, Desel J, Harrer A, Mauser S. Learnflow mining. *DeLFI* 2008, 132:269–280.
37. Schoor C, Bannert M. Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Comput Hum Behav* 2012, 28:1321–1331.
38. Porouhan P, Premchaiswadi W. Process mining and learners' behavior analytics in a collaborative and web-based multi-tabletop environment. *Int J Online Pedagogy Course Design* 2017, 7:29–53.
39. Southavilay V, Yacef K, Callvo RA. Process mining to support students' collaborative writing. In: *Proceedings of the 3th International Conference on Educational Data Mining*, Pittsburgh, USA, 2010:257–266.
40. Cairns AH, Ondo JA, Gueni B, Fhima M, Schwarcfeld M, Joubert C, Khelifa N. Using semantic lifting for improving educational process models discovery and analysis. In: *SIMPDA*. Milan, Italy; 2014, 150–161.
41. Doleck T, Jarrell A, Poitras EG, Chaouachi M, Lajoie SP. Examining diagnosis paths: a process mining approach. In: *Computational Intelligence & Communication Technology (CICT), Second International Conference*. IEEE, Ghaziabad, India, 2016, 663–667.
42. Vahdat M, Oneto L, Anguita D, Funk M, Rauterberg M. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In: *Design for Teaching and Learning in a Networked World*. Toledo, Spain: Springer International Publishing; 2015, 352–366.
43. Ariouat H, Cairns AH, Barkaoui K, Akoka J, Khelifa N. A two-step clustering approach for improving educational process model discovery. In: *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 25th International Conference*. IEEE; Paris, France; 2016, 38–43.
44. Wang R, Zaïane OR. Discovering process in curriculum data to provide recommendation. In: *Proceedings of the 5th International Conference on Educational Data Mining,* Madrid, Spain, 2015, 580–581.
45. Schulte J, Fernandez de Mendonca P, Martinez-Maldonado R, Buckingham Shum S. Large scale predictive process mining and analytics of university degree course data. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM; Vancouver, Canada; 2017, 538–539.
46. Tóth K, Rölke H, Goldhammer F, Barkow I. Educational process mining: new possibilities for understanding students' problem-solving skills. In: *Educational Research and Innovation*. Paris, France: OECD Publishing; 2017, 193–209.
47. Ayutaya NSN, Palungsuntikul P, Premchaiswadi W. Heuristic mining: Adaptive process simplification in education. In: *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 10th International Conference*. IEEE; Bangkok, Thailand; 2012, 221–227.
48. Anuwatvisit S, Tungkasthan A, Premchaiswadi W. Bottleneck mining and Petri net simulation in education situations. In: *ICT and Knowledge Engineering (ICT & Knowledge Engineering), 10th International Conference*. IEEE; Bangkok, Thailand; 2012, 244–251.
49. Poncin W, Serebrenik A, van den Brand M. Process mining software repositories. In: *15th European Conference on Software Maintenance and Reengineering (CSMR)*. IEEE; Burlington, USA; 2011, 5–14.
50. Poncin W, Serebrenik A, van den Brand M. Mining student capstone projects with FRASR and ProM. In: *Proceedings of the ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion*. Portland, USA; 2011, 87–96.
51. Howard L, Johnson J, Neitzel C. Examining learner control in a structured inquiry cycle using process mining. In: *Proceedings of the 3th International Conference on Educational Data Mining*, Pittsburgh, USA, 2010:71–80.
52. Jeong H, Biswas G, Johnson J, Howard L. Analysis of productive learning behaviors in a structured inquiry cycle using hidden Markov models. In: *Proceedings of the 3th International Conference on Educational Data Mining*, Pittsburgh, USA, 2010:81–90.
53. Fernández-Gallego B, Lama M, Vidal JC, Mucientes M. Learning analytics framework for educational virtual worlds. *Proc Comput Sci* 2013, 25:443–447.
54. Sedrakyan G, De Weerdt J, Snoeck M. Process-mining enabled feedback: "tell me what I did wrong" vs. "tell me how to do it right". *Comput Hum Behav* 2016, 57:352–376.
55. Okoye K, Tawil ARH, Naeem U, Lamine E. Discovery and enhancement of learning model analysis through semantic process mining. *Int J Comput Inform Syst Industrial Manage Appl* 2016, 8:93–114.
56. Chen J, Zhang Y, Sun J, Chen Y, Lin F, Jin Q. Personalized micro-learning support based on process mining. In: *7th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE; Huangshan, Anhui, China; 2015, 511–515.
57. Vermeulen M, Mandran N, Labat JM. Chronicle of a scenario graph: from expected to observed learning path. In: *11th European Conference on Technology Enhanced Learning*. Springer Berlin Heidelberg, Lyon, France; 2016, 321–330.

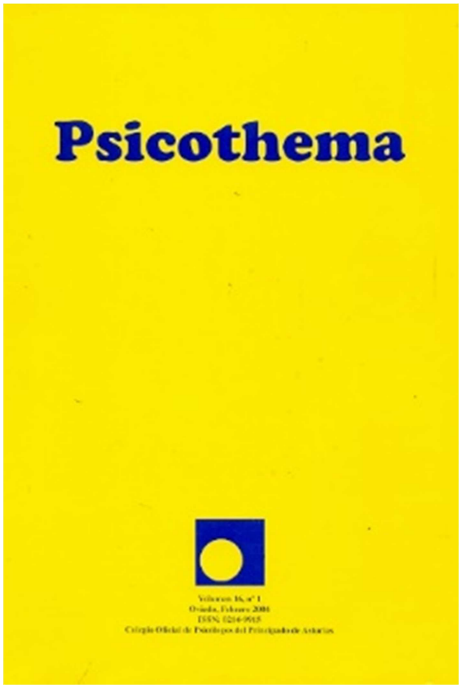
ARTÍCULO 2

Referencia:

- Bogarín, A., Cerezo, R., & Romero, C. (2018). Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs). In: *Psicothema*, *30*(3), 322-329.

Medidas de Calidad Científica:

- Índice de impacto JCR (2017): 1.516
- Área: PSYCHOLOGY, MULTIDISCIPLINARY
- Cuartil en WOS: Q2
- Posición relativa dentro del área: 57 de 135
- Número total de citas en Web of Science: 0
- Número total de citas en GoogleShoolar: 0



Psicothema

Psicothema 2018, Vol. 30, No. 3, 322-329
doi: 10.7334/psicothema2018.116

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs)

Alejandro Bogarín[1], Rebeca Cerezo[2] and Cristóbal Romero[1]
[1] Universidad de Córdoba and [2] Universidad de Oviedo

Abstract

Background: Process mining with educational data has made use of various algorithms for model discovery, principally Alpha Miner, Heuristic Miner, and Evolutionary Tree Miner. In this study we propose the implementation of a new algorithm for educational data called Inductive Miner. **Method:** We used data from the interactions of 101 university students in a course given over one semester on the Moodle 2.0 platform. Data was extracted from the platform's event logs; following preprocessing, the mining was carried out on 21,629 events to discover what models the various algorithms produced and to compare their fitness, precision, simplicity and generalization. **Results:** The Inductive Miner algorithm produced the best results in the tests on this dataset, especially for fitness, which is the most important criterion in terms of model discovery. In addition, when we weighted the various metrics according to their importance, Inductive Miner continued to produce the best results. **Conclusions:** Inductive Miner is a new algorithm which, in addition to producing better results than other algorithms using our dataset, also provides valid models which can be interpreted in educational terms.

Keywords: Educational Data Mining (EDM), Educational Process Mining (EPM), model discovery algorithms, Inductive Miner, Learning Management Systems (LMSs).

Resumen

Descubriendo procesos de aprendizaje aplicando Inductive Miner: un estudio de caso en Learning Management Systems (LMSs). **Antecedentes:** en la minería de procesos con datos educativos se utilizan diferentes algoritmos para descubrir modelos, sobremanera el Alpha Miner, el Heuristic Miner y el Evolutionary Tree Miner. En este trabajo proponemos la implementación de un nuevo algoritmo en datos educativos, el denominado Inductive Miner. **Método:** hemos utilizado datos de interacción de 101 estudiantes universitarios en una asignatura de grado desarrollada en la plataforma Moodle 2.0. Una vez prepocesados se ha realizado la minería de procesos sobre 21.629 eventos para descubrir los modelos que generan los diferentes algoritmos y comparar sus medidas de ajuste, precisión, simplicidad y generalización. **Resultados:** en las pruebas realizadas en nuestro conjunto de datos el algoritmo Inductive Miner es el que obtiene mejores resultados, especialmente para el valor de ajuste, criterio de mayor relevancia en lo que respecta al descubrimiento de modelos. Además, cuando ponderamos con pesos las diferentes métricas seguimos obteniendo la mejor medida general con el Inductive Miner. **Conclusiones:** la implementación de Inductive Miner en datos educativos es una nueva aplicación que, además de obtener mejores resultados que otros algoritmos con nuestro conjunto de datos, proporciona modelos válidos e interpretables en términos educativos.

Palabras clave: minería de datos educativos, minería de procesos educativos, algoritmos de descubrimiento, *Inductive Miner*, sistemas de gestión del aprendizaje.

The increase in internet access in recent years has allowed a huge number of students to experience higher education learning in Computer Based Learning Environments (CBLEs) (Broadbent, & Poon, 2015), generally through widely used Learning Management Systems (LMSs). These systems are ubiquitous in higher education, with 99% of US colleges and universities currently reporting that they have an LMS in place (Dahlstrom, Brooks, & Bichsel, 2014). These LMSs have provided very useful, fairly easy to access information for institutions and stakeholders by collecting data on all student activities at different levels of granularity, from enrollment in a particular program to student performance (Trcka, & Pechenizkiy, 2009). Various educational agents closer to the teaching-learning process have recently started to explore the adoption of these techniques to gain insight into online learners' study processes (Papamitsiou, & Economides, 2014) through Educational Data Mining (EDM), being the first decade of the twenty first century the kick-off of EDM (Peña-Ayala, 2014).

Educational Data Mining (EDM) techniques have been applied extensively to find interesting patterns from large volumes of educational data (Dutt, Ismail, & Herawan, 2017; Romero, & Ventura, 2007). However, EDM techniques are not generally aimed at discovering, analyzing or visualizing the complete educational process, they do not focus on the process but on the result. To allow analysis in which the process plays the central

Received: February 2, 2018 • Accepted: May 9, 2018
Corresponding author: Rebeca Cerezo
Departamento de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: cerezorebeca@uniovi.es

role there is a new line of data-mining research called Educational Process Mining (EPM) (Romero, & Ventura, 2013). Nowadays, the use of Process Mining (PM) in the educational domain is in its early stages and has given rise to EPM research, which is one of the current promising techniques in the EDM firmament (Reimann, Markauskaite, & Bannert, 2014). Although both EDM and EPM start from data, there are some significant differences between them.

EDM can be generally understood as the application of Data Mining (DM) to the specific type of dataset that comes from learning environments in order to address educational questions (Romero, & Ventura, 2010; Weijters, Van Der Aalst, & De Medeiros, 2006). It focuses on the analysis of large data sets in the service of educational science that focuses on modeling and improving learning processes through the use of that data. EPM bridges the gap between EDM and educational science, as it combines data analysis with modeling, and insighting in educational processes. PM is process-centric (Pechenizkiy, Trcka, Vasilyeva, Van Aals &, De Bra, 2009) thereby making unknown (or only partially known) processes explicit. PM, in contrast to DM, is interested in end-to-end processes rather than local patterns (Van der Aalst, 2016).

The goal of EPM is to extract knowledge from event logs recorded by an educational system (LMSs, MOOCs, etc.) and EPM algorithms discover process models of student behavior. There are a great number of PM algorithms for discovering underlying processes from event logs, and they have been used in a wide range of application domains. Most of the work has concentrated on supporting company processes in business contexts (Van Der Aalst, 2011) and although there is a large body of previous research in applying EPM, the algorithms that have been used to report quality metrics to address educational issues are limited to Alpha Miner, Heuristic Miner and Evolutionary Tree Miner (Bogarín, Cerezo, & Romero, 2018):

Alpha Miner (AM): This was the first discovery algorithm and served as the base for the development of later, improved algorithms (Van der Aalst, 2016). Its main limitation is that it doesn't use frequencies, and so does not guarantee soundness, and is only suitable for event logs without noise, quite infrequently fact in learning data.

Heuristic Miner (HM): It has three significant improvements over the Alpha Algorithm. First, it takes frequencies and significance into account, so it can filter out noisy or infrequent behavior, which makes it less sensitive to noise and incomplete logs (Bogarin et al., 2014). Second, it can detect short loops. Third, it allows single activities to be skipped. It does not, however, guarantee sound educational process models.

Evolutionary Tree Miner (ETM): this is a genetic algorithm that optimizes the educational process model based on user-defined quality metrics. In addition, it works with process trees so unsound models will not be considered. By using a genetic algorithm for process discovery, it gains flexibility to change the weighting of different fitness factors, so process discovery can be guided based on the weighted average of predefined quality factors depending on the importance of each factor for the user (Buijs, Van Dongen, & van Der Aalst, 2012).

These algorithms have provided new ways of discovering, monitoring, and improving processes in different educational contexts such as computer-supported collaborative learning, curriculum mining, computer-based assessment, software repositories, professional training, 3D Educational Virtual Worlds, Structured Inquiry Cycle in informal adult learning, and of course, in MOOCs, LMSs and Hypermedia Learning Environments (Bogarin et al., 2018).

Regarding to the LMSs field, Trcka, Pechenizkiy, & Van der Aalst (2010) showed the potential of PM for extracting knowledge from student exam traces in LMSs. In Bogarin, Romero, Cerezo, & Sánchez-Santillán (2014) the authors used data clustering in order to produce more accurate PM models of student behavior. In a similar environment, Reiman et al. in 2014 proposed the use of PM with learning traces based on theoretical principles of Self-Regulated Learning (SRL). Using those principles, Bannert, Reimann, & Sonnenberg (2014) detected differences in frequencies of SRL events using PM techniques. In other research Mukala, Buijs, & Van Der Aalst (2015) used PM techniques in order to trace and analyze successful and unsuccessful student learning patterns based on MOOC data. In later research they also made use of alignment-based conformance checking to analyze students' learning patterns (Mukala, Buijs, Leemans, & Van der Aalst, 2015). Along similar lines, Emond and Buffett (2015) applied process discovery mining and sequence classification mining techniques to model and support SRL in heterogeneous learning environments. Finally, Vidal, Vázquez-Barreiros, Lama, & Mucientes (2016) used logs from a CBLE to extract the learning flow structure using PM, and to obtain the underlying rules that control students' adaptive learning by means of decision tree learning.

Based on current literature, the process discovery algorithm known as *Inductive Miner (IM)* has not been applied to educational datasets until now (Bogarín et al., 2018). In this paper, we propose the use of this algorithm for improving models previously obtained by EPM with other discovery algorithms. Different PM algorithms have been proposed, however no existing algorithm returns good quality metrics in all cases, while IM is being extensively used in business with very promising results (Leemans, Fahland, & van der Aalst, 2013). IM means an improvement over the Alpha and Heuristics miners that makes it easier to explore an event log; it is able to cope with infrequent behavior and large event logs, while ensuring soundness (Leemans, Fahland, & van der Aalst, 2014). It is also expected to produce more sound learning process models. Our objective is to compare the performance of this algorithm with previously used PM algorithms, and the ultimate goal is to be able to produce better process models about student behavior when using CBLEs. Below, we address the study method but describing before the preprocessing data process. Following the results we discuss EPM and its educational value.

Method

Participants

We used data from 101 undergraduate students (mean age=20.23; SD=1.01; female=83%) studying for a degree in psychology at a university in the North of Spain, who completed an online course using the corporate LMS Moodle 2.0.

Instruments

The log file provided by the LMS was the data collection instrument in this study. The data provided by Moodle contains all of each student's events recorded during their interactions with the

LMS, summarized in six attributes (see Table 1). It was necessary to preprocess and filter the Moodle log file; this is essential when we use real event logs and the data is often noisy (Romero, Ventura, & García, 2008).

We converted the students' names into IDs (Identifiers) to maintain their anonymity. Then, we deleted duplicate records, and instructor, system administrator and test user records. We used only four attributes (Time, Full Name, Action and Information) which was sufficient for our research purposes; the name of the course (the same for all records) and the IP address were not relevant. Then, we filtered some irrelevant actions in our log file. So, from the original 42 actions that Moodle stored by default, we selected the 16 actions that were relevant to the learning process and academic performance for this course (Cerezo, Sánchez-Santillán, Paule-Ruiz, & Núñez, 2016). In addition, we used high level coding (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) with five action labels (Planning, Learning, Executing, Review and Forum Peer Learning) in order to produce more easily understandable models (see Table 2) in accordance with assumptions of SRL theory. Following that, we transformed the original Excel log file into the XES (eXtensible Event Stream) file which is required to implement process mining using the ProM framework.

Subsequently, we will consider the student as the "case" and the union between action and high level codification attributes as the "event classes" in only one attribute, for example: URL (Uniform Resource Locator) view-LEARNING, quiz view-PLANNING, and so on. In this way, each row in our preprocessed event logs is an event class (action and high level codification attributes), that is carried out by a case (student) on a specific date (timestamp). The traceability for each case will be the different event classes carried out by a student.

Additionally, we also used the students' final marks. This is a file containing each student's ID and final mark (a numerical value on a 10-point scale). We transformed this continuous value into a categorical value using traditional Spanish academic grading: from 0 to 4.9 is a fail and from 5 to 10 is a pass. Using their performance, we were able to group the students and label them Pass or Fail. Clustering by marks during preprocessing is useful for comparing the performance of different algorithms and for assessing the practical application and theoretical value of the resultant models. In this way we can divide each log file into three different files: All (containing events for all students on the course), Pass (containing only events of students who pass the course) and Fail (containing only events of students who fail the course).

Finally, we produced sub-files by unit in order to analyze student behavior more thoroughly. The course was made up of different units that can be thought of as lessons with different content but similar processes. For this reason we preprocessed the information attribute in each record in order to ascertain which unit it belonged to. Once the preprocessing was done the file was ready for EPM with ProM software (Romero, Cerezo, Bogarín, & Sánchez-Santillán, 2016). Table 3 shows the final number of cases and number of events in each unit after preprocessing.

Table 1
Attributes of a Moodle event log file

Attribute	Description
Course	The name of the course
IP Address	The IP of the device used to access Moodle
Time	The date they accessed Moodle
Full Name	The name of the student
Action	The action that student performed
Information	More information about the action

Table 2
Codification of the attribute actions

Low level Moodle Action	High Level Codification
assign submit	EXECUTING
assign view	PLANNING
forum add discussion	FORUM PEER LEARNING
forum add post	FORUM PEER LEARNING
forum update post	FORUM PEER LEARNING
forum view discussion	FORUM PEER LEARNING
forum view forum	FORUM PEER LEARNING
page view	LEARNING
quiz attempt	EXECUTING
quiz close attempt	EXECUTING
quiz continue attempt	EXECUTING
quiz review	REVIEW
quiz view	PLANNING
quiz view summary	PLANNING
resource view	LEARNING
URL view	LEARNING

Procedure

The experiment took the form of an assignment in the curriculum of a compulsory 3rd year subject completed entirely outside teaching hours. The course was made up of different units that were delivered to the students on a weekly basis during one semester. Students were asked to participate in an eTraining program about SRL and study strategies related to the subject topic (Cerezo, Núñez, Rosario, Valle, Rodríguez, & Bernardo, 2010). The instructor strongly suggested that students approached the assignments for each unit in the following order: understand the theoretical content, put them in practice through the corresponding task, share their experience about the week's topic in the forum; a learning path supported by SRL theory (Nuñez

Table 3
Number of cases and events per unit at the datasets

Units	Number of cases	Number of events
Unit 1	101	1782
Unit 2	101	2103
Unit 3	100	2192
Unit 4	101	2946
Unit 5	100	2514
Unit 6	101	1612
Unit 7	95	2067
Unit 8	87	1931
Unit 9	86	1699
Unit 10	87	1163
Unit 11	84	1620

et al., 2011). However, the students were free to follow their own learning path and the only compulsory assignments for each unit were to complete the weekly practical task and to post at least one comment in each unit forum.

Data analysis

Data analysis had three steps: log file preprocessing (previously described in the *Instruments* section), process discovery, and algorithm evaluation and interpretation (Figure 1).

In order to compare the discovered PM models we executed the most commonly used educational process discovery algorithms provided by the ProM framework (Van der Aalst, 2016): AM algorithm, HM Algorithm, ETM, and finally, the object of this study, *Inductive Miner*. To that end we compared some evaluation measures of the models obtained based on four quality forces (see Figure 2) that measure how well an educational process model describes the observed data:

- *Fitness* quantifies the extent to which the discovered model can accurately reproduce the cases recorded in the log.
- *Precision* shows the proportion of the behavior represented by the model which is not seen in the event log.
- *Generalization* assesses the extent to which the model will be able to reproduce future behavior of the process and can be seen as a measure of confidence in the precision.
- *Simplicity* captures the complexity of a process model in terms of readability.

All indexes are important for process discovery. However, it only makes sense to consider precision, generalization and simplicity if fitness is acceptable (Buijs et al., 2012; Van der Aalst, 2016). Existing process discovery algorithms typically consider, at most, two out of the four main quality dimensions because these four quality forces pull in different directions and whenever one is optimized, quality is usually lost in other measures. In light of this, we used a new *overall* measure proposed by Buijs et al., in 2012, to balance these four measures together, allocating them different weights (see Figure 2) (*Fitness*: weight 10; *Precision*: weight 5; *Generalization*: weight 1; *Simplicity*: weight 1).

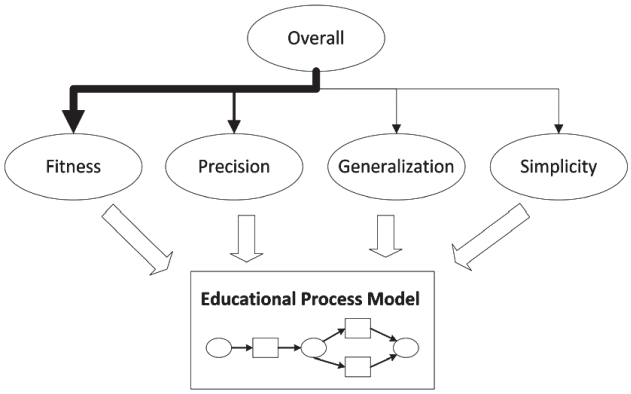


Figure 2. Model quality metrics

Results

Table 4 shows the results of the four algorithms in the *overall* evaluation metric. The IM algorithm scores the highest values in every unit, followed by ETM, then HM and AM depending on the sub-file.

In the *overall* metric, the IM algorithm scored highest, and the same is true if we consider each quality metric separately. Table 5 shows the performance of every algorithm in sub-file 4, where the students showed the most interaction with the LMS resulting in a

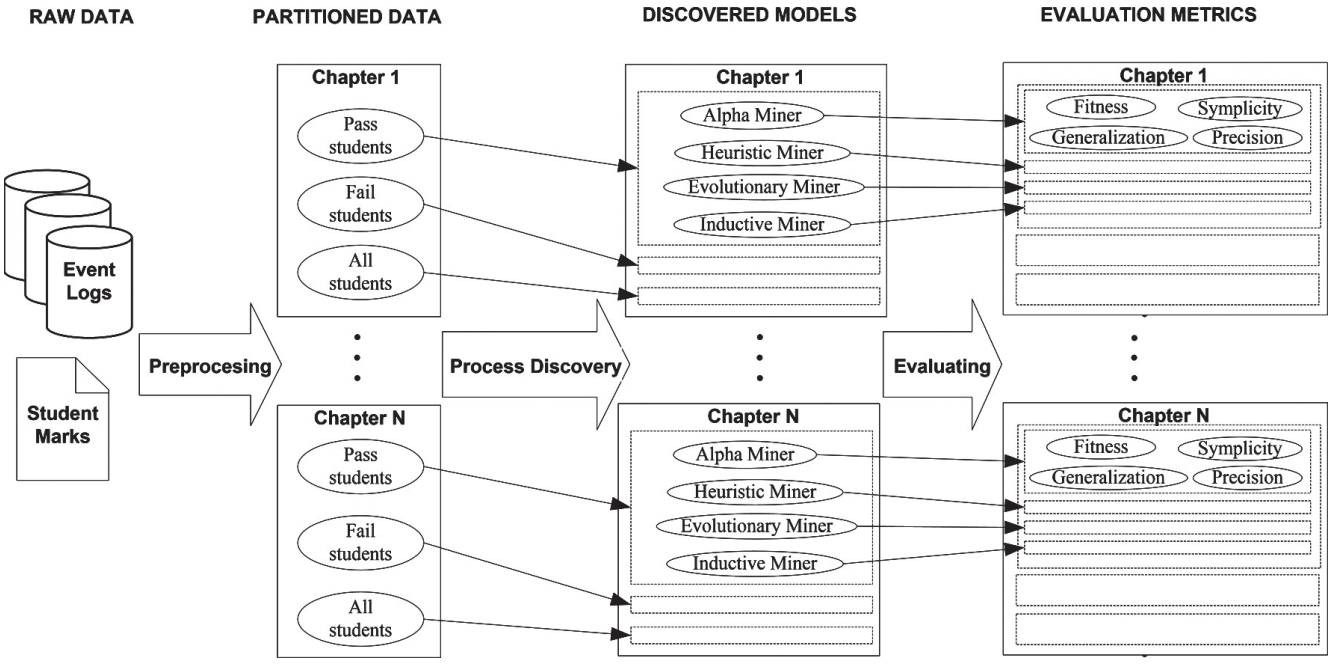


Figure 1. Procedure followed for carrying out EPM

higher number of cases and events, and subsequently complexity of modeling extraction. The IM algorithm scored the highest values in *fitness*, the metric which is fundamental for considering the other quality indexes, and also in *generalization*. It also scored the highest in *simplicity*, *along with* ETM, which indicates that the obtained models are easy to interpret and not *spaghetti-like models*. Nevertheless, it did not achieve the best score for *precision*. The highest scoring models in this metric were ETM and Heuristic Miner, although ETM scored better on *generalization*. The same table also shows the effect of clustering the data, with the quality indexes improving when clustering before, as expected.

Table 4
Comparison of algorithms based on the *overall* metric

Units	AM	HM	ETM	IM
Unit 1	0.676	0.666	0.793	**0.797**
Unit 2	0.666	0.618	0.752	**0.781**
Unit 3	0.583	0.493	0.675	**0.712**
Unit 4	0.452	0.597	0.715	**0.747**
Unit 5	0.582	0.533	0.649	**0.659**
Unit 6	0.577	0.621	0.742	**0.793**
Unit 7	0.612	0.664	0.724	**0.773**
Unit 8	0.724	0.732	0.750	**0.796**
Unit 9	0.516	0.510	0.744	**0.784**
Unit 10	0.685	0.700	0.827	**0.856**
Unit 11	0.553	0.563	0.735	**0.778**

Along with quality metrics, Figures 3 and 4 show two of the resultant visualization of the models for sub-file 4. In order to understand and interpret the IM-generated models it is necessary to understand what each element means: the boxes are the activities carried out by the students, the number in the box is the frequency, the arrows indicate the direction of the process, and the number above the arrows is the frequency of the transition between these two actions. Each model begins with an initial node and ends with a final node.

Looking at the learning path followed by students in the fail cluster (see Figure 3), the first activity they do is the *quiz attempt*, followed by the *quiz view summary* and the *quiz view*. In other words, they start doing activities related to the quiz, which are one of the two compulsory course assignments. In the middle part of the model they do forum-related activities such as *forum view forum* and *forum add post*, which are the other compulsory activities. Following that there are parallel actives -*quiz review, quiz continue attempt*-, finishing with *page view* and *URL view* which would have been the logical starting point for the learning path suggested by the instructor.

Students in the pass cluster (Figure 4) started their study process by visiting the *forum view discussion*, after which the model splits into different possible routes. One route continues

Table 5
Comparison of algorithms based on *fitness, precision, generalization, simplicity*, and *overall* in sub-file 4

Algorithm	Cluster	*Fitness*	*Precision*	*Generalization*	*Simplicity*	*Overall*
Alpha Miner	Fail	0.765	0.197	0.422	0.636	0.570
Heuristic Miner	Fail	0.491	0.521	0.487	0.653	0.509
ET Miner	Fail	0.684	0.709	0.873	0.913	0.716
Inductive Miner	Fail	0.96	0.322	0.957	0.882	**0.768**
Alpha Miner	Pass	0.863	0.164	0.464	0.666	0.622
Heuristic Miner	Pass	0.526	0.707	0.603	0.732	0.596
ET Miner	Pass	0.712	0.691	0.841	0.901	0.725
Inductive Miner	Pass	0.959	0.315	0.962	0.882	**0.765**
Alpha Miner	All	0.581	0.198	0.414	0.466	0.452
Heuristic Miner	All	0.472	0.868	0.483	0.611	0.597
ET Miner	All	0.693	0.715	0.719	0.923	0.715
Inductive Miner	All	0.87	0.443	0.867	0.909	**0.747**

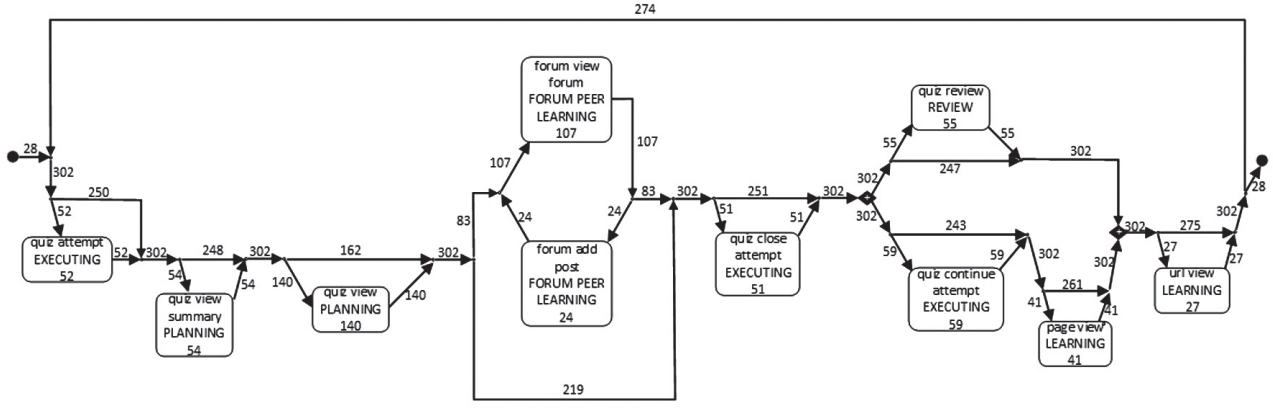


Figure 3. Visualization of failing students' learning path in sub-file 4

via the *URL view* action, the second route involves continuing the study process with forum-related activities -*forum view forum, forum add post* and *forum update post*-. There is also a third route, in which student do actions related to the quizzes *quiz attempt, quiz view summary* and *quiz continue attempt*. The general model finishes with the *quiz close attempt* and *quiz review* actions.

Discussion

This research focuses on the analysis of learning processes based on EPM. We applied PM techniques to educational data in order to discover learning processes, compare algorithm performance, and extract educational implications to guide future work. However, PM algorithms cannot be directly applied to educational problems, preprocessing is necessary first and only then can the mining methods be applied to the problems (Dutt et al., 2017). Therefore, we also described the preprocessing required before discovery could take place.

We proposed the application of the IM algorithm as a new way to discover learning processes in LMSs; an extensive literature review suggested that this is the first study to apply IM to educational data (Bogarin et al., 2018). Based on our results, we can draw three important conclusions. Firstly, the IM algorithm produces the best fitness. This is significant because none of the other quality indexes should be considered in isolation; it only makes sense to consider them all together if the fitness is acceptable (Buijs et al., 2012; van Dongen, 2007). Secondly, the results show that the balance of quality forces (*overall*) are better in IM than in the other contemporary PM algorithms. And thirdly, both metrics, taken together or individually, are even better when we apply clustering to improve subsequent mining, as previously seen with educational (Bogarin et al., 2014; Bogarin et al., 2018) and business data (Bose, & van der Aalst, 2009). It seems that, applying the IM algorithm to discovering learning models opens a new field in the research, development, and understanding of PM applied to educational issues.

Process discovery is one of the most challenging process mining tasks; starting from a simple log, a process model is constructed capturing the behavior seen in the log (Van der Aalst, 2011). However, apart from quality compliance, the resultant model needs to be able to reproduce the behavior seen in the log file in an understandable educational process, giving EPM a practical meaning rather than just ideas and theories.

With that in mind, we selected and interpreted two of the most challenging discovered models. If, in addition to the raw actions in the failing cluster, we look at the high level coding, the models lead us to conclude that the students who failed did not follow the learning path suggested by the instructor and promoted by SRL theories leading to quality learning results. Based on the assumptions of SRL (Zimmerman, 1990), starting executing before planning or learning leads to low quality learning or failure, as seen in this study.

If we look at the high-level coding of those students who passed, we can see that although they did not follow the instructors' suggestions exactly, they did follow the logic of a successful learning process. These results are in line with those from Lust, Elen, and Clarebout (2013a, 2013b), who found that only a minority of students regulated their behavior in line with course requirements. Passing students started with actions indicating comprehension and learning of the materials, three different routes can then be seen: two task oriented groups, one socially focused giving a leading role to collaborative learning in the forums, and another more individually focused; and finally, a non-task or learning oriented group. These different learning profiles are in accordance with data previously obtained by Cerezo et al, 2016 also using LMS interaction data. All three routes concluded with the executing and reviewing actions suggested by the instructor and SRL rationale, leading to successful achievement in varying degrees.

The PM models also allow us to examine which specific actions the students performed. It is interesting to see the actions related to forum-supported collaborative learning. Students in

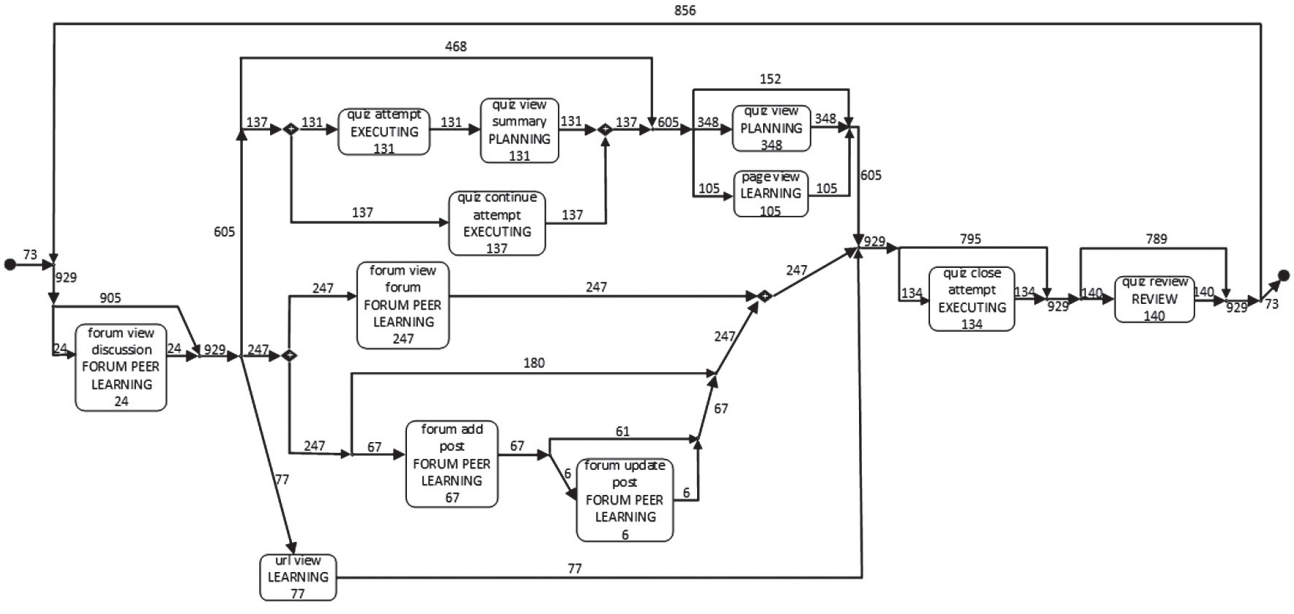


Figure 4. Visualization of passing students' learning path in sub-file 4

the pass cluster performed actions such as *forum update post* and *forum view discussion*, which do not appear in the model from the fail cluster. This is very valuable since forum behavior has been previously related to student achievement in LMSs (Romero, López, Luna & Ventura, 2013).

It should be also noted that IM models are able to discover meaningful learning processes similar to those previously obtained using alternative algorithms (Mukala, Buijs, & Van Der Aalst, 2015; Mukala, Buijs, et al., 2015). However, in this case the instructor can visualize and interpret the behavior model of students' learning paths thanks to their simplicity. Process discovery algorithms often result in spaghetti-like process models (Van der Aalst, 2011), which are very hard to read. However, IM strongly focuses on simplicity and generally results in simple models (Buijs et el., 2012).

In conclusion, learning model discovery with IM could be a promising resource for preventing learning failure in LMSs. With insight into at-risk students' distance-learning progress, we can strategically design preventive interventions based on Adaptive Hypermedia Learning Environments (Brusilovsky, & Millán, 2007) or early detection and remedial actions through real time modeling. PM is not restricted to the past, but also relevant to the present (recommendation and real-time conformance checking), and the future (prediction) (Van der Aalst, Schonenberg, & Song, 2011). In a wider sense, the scope in academic contexts is also extensive, from allow universities to invest in those resources which are shown to be most useful for preventing school drop-out (Areces, Rodríguez Muñiz, Suárez Álvarez, de la Roca, & Cueli, 2016) to the contribution of social networks to learning (Sanmamed, Carril, & Alvarez de Sotomayor, 2017).

Finally, in order to generalize the good performance of IM with educational data, it would be interesting to test the algorithm in different CBLEs, such as alternative LMSs or the emerging MOOCs. Modeling learning process in MOOCs would be a very challenging prospect in terms of simplicity and readability.

Acknowledgements

Authors gratefully acknowledge the financial subsidy provided by Spanish Ministry of Science and Technology TIN2017-83445-P and EDU2014-57571-P. We have also received funds from the European Union and the Principality of Asturias, through its Science, Technology and Innovation Plan (grant GRUPIN14-053).

References

Areces, D., Rodríguez Muñiz, L. J., Suárez Álvarez, J., de la Roca, Y., & Cueli, M. (2016). Information sources used by high school students in the college degree choice. *Psicothema*, *28*(3), 253-259. doi: 10.7334/psicothema2016.76

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. In: *Metacognition and learning*, *9*(2), 161-185. doi:10.1007/s11409-013-9107-6

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In M. Pistilli, J. Willis, & D. Koch (Eds.), *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 170-181). Indianapolis, USA: ACM. doi:10.1145/2567574.2567604

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1). doi:10.1002/widm.1230

Bose, R. J. C., & van der Aalst, W. M. (2009, September). Trace clustering based on conserved patterns: Towards achieving better process models. In U. Dayal, J. Eder, J. Koehler & H. Reijers (Eds.), *Proceedings of the International Conference on Business Process Management* (pp. 170-181). Berlin, Heidelberg: Springer.

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. doi:10.1016/j.iheduc.2015.04.007

Buijs, J. C., Van Dongen, B. F., & van Der Aalst, W. M. (2012). On the role of fitness, precision, generalization and simplicity in process discovery. In R. Meersman, H. Panetto, T. Dillon, S. Rinderle-Ma, P, Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, & I. F. Cruz, *Proceedings of the OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 305-322). Berlin: Springer. doi:10.1007/978-3-642-33606-5_19

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovski, A. Kobsa & W. Nejdl (Eds.), *The adaptive web* (pp. 3-53). Berlin: Springer.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42-54. doi:10.1016/j.compedu.2016.02.006

Cerezo, R., Núñez, J. C., Rosario, P., Valle, A., Rodríguez, S., & Bernardo, A. (2010). New Media for the promotion of self-regulated learning in higher education. Psicothema, *22*(2), 306-315.

Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). *The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives* (Research report) Retrieved from http://www. educause. edu/ecar. 2014 EDUCAUSE. CC by-nc-nd

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005. doi:10.1109/ACCESS.2017.2654247

Emond, B., & Buffett, S. (2015, June). *Analyzing Student Inquiry Data Using Process Discovery and Sequence Classification*. Paper presented at the International Educational Data Mining Society, Madrid, Spain.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34. doi: 10.1145/240455.240464

Leemans, S. J., Fahland, D., & van der Aalst, W. M. (2013, August). *Discovering block-structured process models from event logs containing infrequent behaviour*. Paper presented at the International Conference on Business Process Management, Beijing, China.

Leemans, S. J., Fahland, D., & van der Aalst, W. M. (2014). Process and Deviation Exploration with Inductive Visual Miner. *BPM (Demos)*, *1295*, 46.

Lust, G., Elen, J., & Clarebout, G. (2013a). Regulation of tool-use within a blended course: student differences and performance effects. *Computers & Education*, *60*(1), 385-395.

Lust, G., Elen, J., & Clarebout, G. (2013b, August). *Measuring students' strategy-use within a CMS supported course through students' tool-use patterns*. Paper presented at the 15th biennial conference EARLI 2013, Munich, Germany.

Mukala, P., Buijs, J. C. A. M., & Van Der Aalst, W. M. P. (2015). *Uncovering learning patterns in a MOOC through conformance alignments* (Research report). Retrieved from http://bpmcenter.org/wp-content/uploads/reports/2015/BPM-15-09.pdf

Mukala, P., Buijs, J. C., Leemans, M., & van der Aalst, W. M. (2015, December). *Learning Analytics on Coursera Event Data: A Process Mining Approach*. Paper presented at the SIMPDA, Viena, Austria.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review

of empirical evidence. *Journal of Educational Technology & Society, 17*(4), 49.

Pechenizkiy, M., Trcka, N., Vasilyeva, E., van Aalst, W., & De Bra, P. (2009, July). *Process mining online assessment data*. In *Educational Data Mining*. Paper presented at the International Conference on Educational Data Mining, Córdoba, Spain.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462. doi:10.1016/j.eswa.2013.08.042

Reimann, P., Markauskaite, L., & Bannert, M. (2014). E-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, *45*(3), 528-540. doi:10.1111/bjet.12146

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146. doi:10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, 40*(6), 601-618. doi:10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384. doi: 10.1016/j.compedu.2007.05.016

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: a tutorial and case study using Moodle data sets. In *Data Mining and Learning Analytics: Applications in Educational Research* (pp. 1-28). Wiley & Blackwell. doi:10.1002/9781118998205.ch1

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68,* 458-472.

Sanmamed, M. G., Carril, P. C. M., & Álvarez De Sotomayor, I. D. (2017). Factors which motivate the use of social networks by students. *Psicothema, 29*(2), 204-210. doi: 10.7334/psicothema2016.127

Trcka, N., & Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications* (pp. 1114-1119). New Jersey: The Institute of Electrical and Electronics Engineers. doi:10.1109/ISDA.2009.159

Trcka, N., Pechenizkiy, M., & van der Aalst, W. (2010). Process mining from educational data. In C. Romero, S. Ventura, M. Pechenizkiy & R. Baker (Eds.), *Handbook of educational data mining* (pp. 123-142). Florida: Taylor & Francis.

van der Aalst, W. M. (2011). Process Discovery: An Introduction. In *Process Mining* (pp. 125-156). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-19345-3_5

van der Aalst, W. M. (2016). *Process mining: data science in action*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-49851-4

van der Aalst, W. M., Schonenberg, M. H., & Song, M. (2011). Time prediction based on process mining. *Information systems*, *36*(2), 450-475.

van Dongen, B. F. (2007). Process mining and verification. *Dissertation Abstracts International*, *68*(4).

Vidal, J. C., Vázquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems, 100,* 160-174. doi:10.1016/j.knosys.2016.03.003

Weijters, A.J.M.M., van Der Aalst, W.M., & De Medeiros, A.A. (2006). Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven Technology Reports, 166*, 1-34.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational psychologist, 25*(1), 3-17.

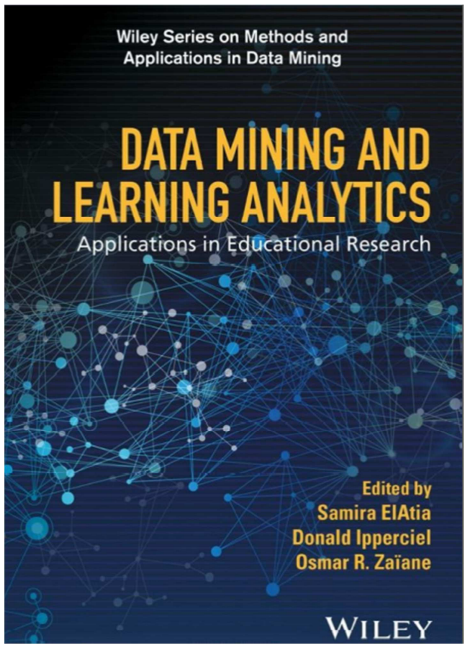
ARTÍCULO 3

Referencia:

- Romero, C., Cerezo, R., Bogarín, A., & Sánchez□Santillán, M. (2016). Educational process mining: A tutorial and case study using moodle data sets. In: *ElAtia, S., Ipperciel, D., & Zaïane, O.R. (Eds.). Data Mining and Learning Analytics: Applications in Educational Research*. 3-28.

Medidas de Calidad Científica:

- Índice de impacto: 1.26
- Área: ENGINEERING AND TECHNOLOGY
- Cuartil: Q1
- Posición relativa dentro del área: 7 de 37
- Número total de citas en Web of Science: 5
- Número total de citas en GoogleShoolar: 7



PART I

AT THE INTERSECTION OF TWO FIELDS: EDM

CHAPTER 1

EDUCATIONAL PROCESS MINING: A TUTORIAL AND CASE STUDY USING MOODLE DATA SETS

Cristóbal Romero[1], Rebeca Cerezo[2], Alejandro Bogarín[1], and Miguel Sánchez-Santillán[2]

[1] Department of Computer Science, University of Córdoba, Córdoba, Spain
[2] Department of Psychology, University of Oviedo, Oviedo, Spain

The use of learning management systems (LMSs) has grown exponentially in recent years, which has had a strong effect on educational research. An LMS stores all students' activities and interactions in files and databases at a very low level of granularity (Romero, Ventura, & García, 2008). All this information can be analyzed in order to provide relevant knowledge for all stakeholders involved in the teaching–learning process (students, teachers, institutions, researchers, etc.). To do this, data mining (DM) can be used to extract information from a data set and transform it into an understandable structure for further use. In fact, one of the challenges that the DM research community faces is determining how to allow professionals, apart from computer scientists, to take advantage of this methodology. Nowadays, DM techniques are applied successfully in many areas, such as business marketing, bioinformatics, and education. In particular, the area that applies DM techniques in educational settings is called educational data mining (EDM). EDM deals with unintelligible, raw educational data, but one of the core goals of this discipline—and the present chapter—is to make this valuable data legible and usable to students as feedback, to professors as assessment, or to universities for strategy. EDM is broadly studied, and a reference tutorial was developed by Romero et al. (2008). In this tutorial, the authors show the step-by-step process for doing DM with Moodle data. They describe how to apply preprocessing and traditional DM techniques

Data Mining and Learning Analytics: Applications in Educational Research, First Edition.
Edited by Samira ElAtia, Donald Ipperciel, and Osmar R. Zaïane.

(such as statistics, visualization, classification, clustering, and association rule mining) to LMS data.

One of the techniques used in EDM is process mining (PM). PM starts from data but is process centric; it assumes a different type of data: events. PM is able to extract knowledge of the event log that is commonly available in current information systems. This technique provides new means to discover, monitor, and improve processes in a variety of application domains. The implementation of PM activities results in models of business processes and historical information (more frequent paths, activities less frequently performed, etc.). Educational process mining (EPM) involves the analysis and discovery of processes and flows in the event logs generated by educational environments. EPM aims to build complete and compact educational process models that are able to reproduce all the observed behaviors, check to see if the modeling behavior matches the behavior observed, and project extracted information from the registrations in the pattern to make the tacit knowledge explicit and to facilitate a better understanding of the process (Trcka & Pechenizkiy, 2009).

EPM has been previously applied successfully to the educational field; one of the most promising applications is used to study the difficulties that students of different ages show when learning in highly cognitively and metacognitively demanding learning environments, such as a hypermedia learning environment (Azevedo et al., 2012). These studies describe suppositions and commonalities across several of the foremost EPM models for self-regulated learning (SRL) with student-centered learning environments (SCLEs). It supplies examples and definitions of the key metacognitive monitoring processes and the regulatory skills used when learning with SCLEs. It also explains the assumptions and components of a leading information processing model of SRL and provides specific examples of how EPM models of metacognition and SRL are embodied in four current SCLEs.

However, several problems have been previously found when using EPM (Bogarín et al., 2014). For instance, the model obtained is not well adjusted to the general behavior of students, and the resulting model may be too large and complex for a teacher or student to analyze. In order to solve these problems, we propose the use of clustering for preprocessing the data before applying EPM to improve understanding of the obtained models. Clustering techniques divide complex phenomena—described by sets of objects or by highly dimensional data—into small, comprehensible groups that allow better control and understanding of information. In this work, we apply clustering as a preprocessing task for grouping users based on their type of course interactions. Thus, we expect to discover the most specific browsing behaviors when using only the clustered data rather than the full data set. This chapter describes, in a practical tutorial, how to apply clustering and EPM to Moodle data using two well-known open-source tools: Weka (Witten, Frank, & Hall, 2011) and ProM (Van der Aalst, 2011a).

The chapter is organized as follows: Section 1.1 describes the most relevant works related to the chapter, Section 1.2 describes the data preparation and clustering, Section 1.3 describes the application of PM, and Section 1.4 outlines some conclusions and suggestions for further research.

1.1 BACKGROUND

Process mining (PM) is a data mining (DM) technique that uses event logs recorded by systems in order to discover, monitor, and improve processes in different domains. PM is focused on processes, but it also uses the real data (Van der Aalst, 2011a). It is the missing link between the classical process model of analysis and data-oriented analysis like DM and machine learning. We can think of PM as a bridge between processes and data, between business process management and business intelligence, and between compliance and performance. PM connects many different ideas, and that makes it extremely valuable (Van der Aalst, 2011b).

The starting point for PM is event data. We assume that there is an event log in which each event refers to a case, an activity, and a point in time or time stamp. An event log can be seen as a collection of cases (which we sometimes also refer to as traces); each case corresponds to a sequence of events. Event data comes from a large variety of sources. PM consists of different types of mining (Van der Aalst et al., 2012):

- **Process discovery** conforms to a model.
- **Conformance checking** is a form of replay aimed at finding deviations.
- **Enhancement** is also a form of replay with the goal of finding problems (such as bottlenecks) or ideas for improvement.

The potential and challenges of PM have been previously investigated in the field of professional training (Cairns et al., 2014). For instance, this field has focused on the mining and analysis of social networks involving course units or training providers; it has also proposed a two-step clustering approach for partitioning educational processes following key performance indicators. Sedrakyan, Snoeck, and De Weerdt (2014) attempted to obtain empirically validated results for conceptual modeling of observations of activities in an educational context. They tried to observe the characteristics of the modeling process itself, which can be associated with better/worse learning outcomes. In addition, the study provided the first insights for learning analytics research in the domain of conceptual modeling.

The purpose of another interesting study, which was conducted by Schoor and Bannert (2012), was to explore sequences of social regulatory processes during a computer-supported collaborative learning task and to determine these processes' relationship to group performance. Using an analogy to self-regulation during individual learning, the study conceptualized social regulation as both individual and collaborative activities: analyzing, planning, monitoring, and evaluating cognitive and motivational aspects during collaborative learning. In an exploratory way, the study used PM to identify process patterns for high and low group performance dyads.

Referring to the research on self-regulated learning (SRL), the recent work of Bannert, Reimann, and Sonnenberg (2014) analyzed individual regulation in terms of a set of specific sequences of regulatory activities. Thus, the aim of the study's approach was to analyze the temporal order of spontaneous individual regulation activities. This research demonstrates how various methods developed in the PM

research can be applied to identify process patterns in SRL events, as captured in verbal protocols. It also shows how theoretical SRL process models can be tested with PM methods.

Another related work observed how linking labels in event logs to their underlying semantics can bring educational process discovery to the conceptual level (Cairns et al., 2004). In this way, more accurate and compact educational processes can be mined and analyzed at different levels of abstraction. It is important to say that this approach was done using the ProM framework (Van der Aalst, 2011a).

ProM contains the Heuristics Miner plug-in, which has been used to analyze a student's written activities and thus to improve the student's writing skills. Southavilay, Yacef, and Calvo (2010) presented a job that enables the development of a basic heuristic to extract the semantic meaning of text changes and determine writing activities. Heuristics have been able to analyze the activities of student writing using PM and have found patterns in these activities. The discovered patterns, the snapshot of processes provided by the sequence of action, and the dotted chart analysis can be used to provide feedback to students so that they are aware of their writing activities. One way to improve understanding of how writing processes lead to a better outcome is to improve heuristics (Boiarsky, 1984). In this work, only changes in spelling, numbers, ratings, and formats are considered. No grammatical corrections are included. In addition, one of the proposed changes is vocabulary improvement. Another concept that is not taken into account in this work is the repetition of words; good writers often avoid the annoying repetition of words and instead use synonyms. Finally, the Heuristic Miner was previously used to investigate the processes recorded by students at the University of Thailand to minimize the educational adaptation process (Ayutaya, Palungsuntikul, & Premchaiswadi, 2012). The referenced work demonstrated the behavior of Heuristic Miner in the extraction of a slightly structured process. The properties of the Heuristics Miner plug-in were shown using an event log from the University of Thailand. In addition, the Heuristics Miner was also used to analyze learning management system (LMS) learning routes and to track the behavior learned in relation to the respective learning styles, which must be identified in advance.

On the other hand, the process of grouping students is also very relevant for educational data mining (EDM). This naturally refers to an area of data analysis, namely, data clustering, which aims to discover the natural grouping structure of a data set. A pair of good reviews was conducted about the application of clustering techniques for improving e-learning environments. The first review, by Vellido, Castro, and Nebot (2011), was devoted to clustering educational data and its corresponding analytical methods; these methods hold the promise of providing useful knowledge to the community of e-learning practitioners. The authors of this review described clustering and visualization methods that enhance the e-learning experience due to the capacity of the former to group similar actors based on similarities and the ability of the latter to describe and explore these groups intuitively.

The second review, by Dutt et al. (2015), aimed to consolidate the different types of clustering algorithms applied in the EDM context and to answer the question of how a higher educational institution can harness the power of didactic data for strategic use. Building an information system that can learn from data is a difficult

task, but it has been achieved using various data mining approaches, such as clustering, classification, and prediction algorithms.

Finally, we want to note that we found no works that use clustering techniques together with educational process mining (EPM). Thus, with the exception of our previous works (Bogarín et al., 2014), to our knowledge, there have been no published works about this topic. In fact, this chapter is an in-depth extension of our previous short work (Bogarín et al., 2014); we have reoriented this chapter to be a practical guide that can also be used by a nonexpert, such as an instructor.

1.2 DATA DESCRIPTION AND PREPARATION

The data sets used in this work were gathered from a Moodle 2.0 course used by 84 undergraduate students from the psychology degree program at a university in northern Spain. The experiment was implemented during two semesters as an assignment for a third-year compulsory subject. Students were asked to participate in an e-learning/training program about "learning to learn" and a SRL that was to be completed entirely outside of teaching hours. The program was made up of 11 different units that were sent to the students on a weekly basis, and each student was able to work on each unit for a 15-day period. Students got an extra point on their final subject grade if they completed at least 80% of the assignments.

1.2.1 Preprocessing Log Data

Moodle logs every click that students make for navigational purposes (Van der Aalst et al., 2012). Moodle has a modest built-in log-viewing system (see Fig. 1.1). Log files can be filtered by course, participant, day, and activity, and they can be shown

Course	IP Address	Date	Full name	Action	Information
Trastornos del Apre	156.35.71.136	2-10-2012-12:35	FERNANDEZ MARTINEZ Carla	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	81.9.215.5	2-10-2012-12:58	ALVAREZ SAN MILLAN Andrea	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	81.9.215.5	2-10-2012-12:58	ALVAREZ SAN MILLAN Andrea	questionnaire view	PROYECTO E-TRAL
Trastornos del Apre	156.35.221.243	2-10-2012-13:30	HOMBACH VIOLA MARIANNA	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	page view	Hoja de Ruta
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	folder view all	
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	resource view all	
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	label view all	
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	page view all	
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	imscp view all	
Trastornos del Apre	156.35.221.243	2-10-2012-13:32	HOMBACH VIOLA MARIANNA	url view all	
Trastornos del Apre	83.97.248.62	2-10-2012-13:51	CARRIO CARRO Luis	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	93.156.24.124	2-10-2012-14:16	Rodríguez Carballo Andrea	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	156.35.221.243	2-10-2012-14:23	HOMBACH VIOLA MARIANNA	page view	Carta cero
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	label view all	
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	imscp view all	
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	resource view all	
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	url view all	
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	page view all	
Trastornos del Apre	156.35.221.243	2-10-2012-15:05	HOMBACH VIOLA MARIANNA	folder view all	
Trastornos del Apre	80.39.86.208	2-10-2012-15:16	Sánchez Sánchez María	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	88.29.14.156	2-10-2012-15:23	García Pérez LAURA	course view	Trastornos del Aprendizaje (Grado en Psicología)
Trastornos del Apre	88.29.14.156	2-10-2012-15:24	García Pérez LAURA	forum view forum	Novedades
Trastornos del Apre	88.29.14.156	2-10-2012-15:24	García Pérez LAURA	forum view forum	Tablón de Anuncios
Trastornos del Apre	88.29.14.156	2-10-2012-15:24	García Pérez LAURA	page view	Hoja de Ruta
Trastornos del Apre	88.29.14.156	2-10-2012-15:25	García Pérez LAURA	page view	Carta cero
Trastornos del Apre	156.35.71.136	2-10-2012-15:34	Alonso Vega Jesus	course view	Trastornos del Aprendizaje (Grado en Psicología)

Figure 1.1 Moodle event log.

or saved in files with the formats: text format (TXT), open document format for office applications (ODS), or Microsoft excel file format (XLS).

We did not use all the information included in the Moodle log file provided (see Table 1.1). In particular, we did not use the name of the course (because it is the same for all records) or the internet protocol (IP) address (because it is irrelevant for our purposes).

Additionally, we have also filtered the log file to eliminate those records that contain an action that could be considered irrelevant to the students' performance. Thus, from all the actions that Moodle stored in our log file (39 in total), we only used the 20 actions that were related to the students' activities in the course (see Table 1.2). This filter lets us reduce the log file from 41,532 to 40,466 records.

Then, we created a new attribute by joining the action and information attributes. We implemented this transformation because it provides additional valuable

TABLE 1.1 Variables of the Moodle log file

Attribute	Description
Course	The name of the course
IP address	The IP of the device used to access
Time	The date they accessed it
Full name	The name of the student
Action	The action that the student has done
Information	More information about the action

TABLE 1.2 Actions considered relevant to the students' performance

assignment upload
assignment view
course view
folder view
forum add discussion
forum add post
forum update post
forum view discussion
forum view forum
page view
questionnaire submit
questionnaire view
quiz attempt
quiz close attempt
quiz continue attempt
quiz review
quiz view
quiz view summary
resource view
url view

TABLE 1.3 List of events in the quiz view after joining action and information

quiz view: Actividad 11
quiz view: Actividad 4
quiz view: Actividad 6
quiz view: Actividad 7
quiz view: Actividad 9
quiz view: Actividad Mapa Conceptual
quiz view: Actividad Tema 2
quiz view: Actividad Tema 3 El Código Secreto
quiz view: Actividad Tema 3 Toma de apuntes
quiz view: Carta 1
quiz view: Carta 10
quiz view: Carta 11
quiz view: Carta 2
quiz view: Carta 3
quiz view: Carta 4
quiz view: Carta 5
quiz view: Carta 6
quiz view: Carta 7
quiz view: Carta 8
quiz view: Carta 9
quiz view: Neutra II
quiz view: Neutra III
quiz view: Subrayado y resumen
quiz view: Tarea neutra enfermedad
quiz view: Tarea: Aprende a Relajarte

information related to the action. For example, a particular action in the quiz view was associated with 25 different information fields, as shown in Table 1.3 (action: information). After completing this transformation, we obtained a total of 332 events (actions plus the information field) that students executed when browsing the course.

Finally, it was necessary to transform the files into the appropriate format for use by the ProM (Van der Aalst, 2011a) tool. To do this, the Moodle log file was firstly saved in the comma-separated values (CSV) format, as shown in Figure 1.2.

Then, the CSV file was converted to mining extensible markup language (MXML), which is the format interpreted by ProM. We used the ProM Import Framework to do this conversion. We selected the option "General CSV File" from the "Filter" properties tab (see Fig. 1.3), and we linked the names of the head of this CSV file with corresponding labels in the properties panel:

- The "Case ID" property was linked with the "Action" value.
- The "Task ID" property was linked with the "Information" value.
- The "Start Time" property was linked with the "Time" value.
- The "Originator" property was linked with the "Full Name" value.

It is also important to set the "Date Format" field correctly; in this case, the format is "D-M-Y-H: M."

total.csv

```
Time;FullName;Action;Information
10-10-2012-19:28;Pisatti Combina Santiago Matias;course view;course view
10-10-2012-19:28;Pisatti Combina Santiago Matias;forum view discussion;forum view discussion
10-10-2012-19:28;Pisatti Combina Santiago Matias;forum view forum;forum view forum
10-10-2012-19:40;Pisatti Combina Santiago Matias;forum view discussion;forum view discussion
10-11-2012-18:26;Pisatti Combina Santiago Matias;course view;course view
10-11-2012-18:26;Pisatti Combina Santiago Matias;quiz attempt;quiz attempt
10-11-2012-18:26;Pisatti Combina Santiago Matias;quiz continue attempt;quiz continue attempt
10-11-2012-18:26;Pisatti Combina Santiago Matias;quiz view;quiz view
10-11-2012-18:32;Pisatti Combina Santiago Matias;course view;course view
10-11-2012-18:32;Pisatti Combina Santiago Matias;quiz close attempt;quiz close attempt
10-11-2012-18:32;Pisatti Combina Santiago Matias;quiz review;quiz review
10-11-2012-18:32;Pisatti Combina Santiago Matias;quiz view summary;quiz view summary
10-11-2012-18:32;Pisatti Combina Santiago Matias;resource view;resource view
10-1-2013-16:26;Pisatti Combina Santiago Matias;course view;course view
10-1-2013-16:27;Pisatti Combina Santiago Matias;folder view;folder view
11-10-2012-19:55;Pisatti Combina Santiago Matias;course view;course view
11-10-2012-19:56;Pisatti Combina Santiago Matias;forum view discussion;forum view discussion
11-10-2012-19:56;Pisatti Combina Santiago Matias;forum view forum;forum view forum
11-10-2012-19:58;Pisatti Combina Santiago Matias;forum view discussion;forum view discussion
1-11-2012-20:34;Pisatti Combina Santiago Matias;course view;course view
1-11-2012-20:34;Pisatti Combina Santiago Matias;forum view discussion;forum view discussion
1-11-2012-20:34;Pisatti Combina Santiago Matias;forum view forum;forum view forum
11-1-2013-19:58;Pisatti Combina Santiago Matias;course view;course view
11-1-2013-19:58;Pisatti Combina Santiago Matias;questionnaire view;questionnaire view
1-1-2013-18:31;Pisatti Combina Santiago Matias;course view;course view
12-11-2012-16:04;Pisatti Combina Santiago Matias;course view;course view
12-11-2012-16:04;Pisatti Combina Santiago Matias;quiz view;quiz view
12-11-2012-20:41;Pisatti Combina Santiago Matias;course view;course view
```

Figure 1.2 Moodle event log in CSV format.

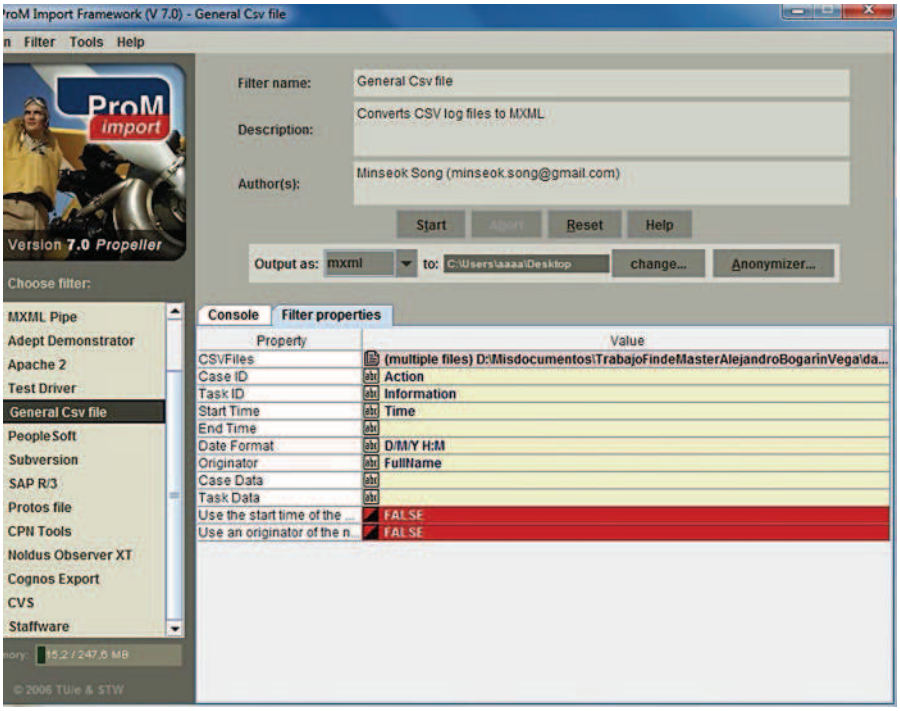


Figure 1.3 Interface for the ProM import tool.

total.mxml

```
<?xml version="1.0" encoding="UTF-8" ?>
<!-- MXML version 1.1 -->
<!-- Created by ProM Import Framework, Version 7.0 (Propeller) -->
<!-- via MXMLib Version 1.9 (http://promimport.sf.net/) -->
<!-- (c) 2004-2007 C.W. Guenther (christian@deckfour.org); Eindhoven Technical University -->
<!-- This event log is formatted in MXML, for use by BPI and Process Mining Tools. -->
<!-- You can load this file e.g. in the ProM Framework for Process Mining. -->
<!-- More information about MXML, Process Mining, and ProM: http://www.processmining.org/. -->
<WorkflowLog xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="http://is.tm.tue.nl/research/processmining/WorkflowLog.xsd"
description="Unified single process">
    <Data>
        <Attribute name="app.name">ProM Import Framework</Attribute>
        <Attribute name="app.version">7.0 (Propeller)</Attribute>
        <Attribute name="java.vendor">Oracle Corporation</Attribute>
        <Attribute name="java.version">1.7.0_21</Attribute>
        <Attribute name="mxml.creator">MXMLib (http://promimport.sf.net/)</Attribute>
        <Attribute name="mxml.version">1.1</Attribute>
        <Attribute name="os.arch">x86</Attribute>
        <Attribute name="os.name">Windows XP</Attribute>
        <Attribute name="os.version">5.1</Attribute>
        <Attribute name="user.name">alex</Attribute>
    </Data>
    <Source program="CSV files"/>
    <Process id="UNIFIED" description="Unified single process">
        <ProcessInstance id="assignment view">
            <AuditTrailEntry>
                <WorkflowModelElement>assignment view</WorkflowModelElement>
                <EventType>start</EventType>
                <Timestamp>2012-10-15T09:00:00.000+02:00</Timestamp>
                <Originator>HOMBACH VIOLA MARIANNA</Originator>
            </AuditTrailEntry>
            <AuditTrailEntry>
                <WorkflowModelElement>assignment view</WorkflowModelElement>
                <EventType>complete</EventType>
                <Timestamp>2012-10-15T09:00:00.000+02:00</Timestamp>
                <Originator>HOMBACH VIOLA MARIANNA</Originator>
            </AuditTrailEntry>
```

Figure 1.4 MXML file for use with ProM.

The file resulting from the filter is shown in Figure 1.4. This file is then used with ProM in order to do EPM.

1.2.2 Clustering Approach for Grouping Log Data

We also propose an approach for using clustering as a preprocessing task for improving EPM. The traditional approach uses all event log data to disclose a process model of a student's behavior. However, this approach applies clustering first in order to group students with similar marks or characteristics; then, it implements PM to discover more specific models of the student's behavior (see Fig. 1.5).

The proposed approach used two clustering/grouping methods:

1. *Manual clustering*: grouping students directly using only the students' marks on the course's final exam.
2. *Automatic clustering*: grouping students using a clustering algorithm based on their interactions with the Moodle course.

Manual clustering uses the student's final mark, which is a numeric value on a 10-point scale provided by the instructor. We turned this continuous value into a

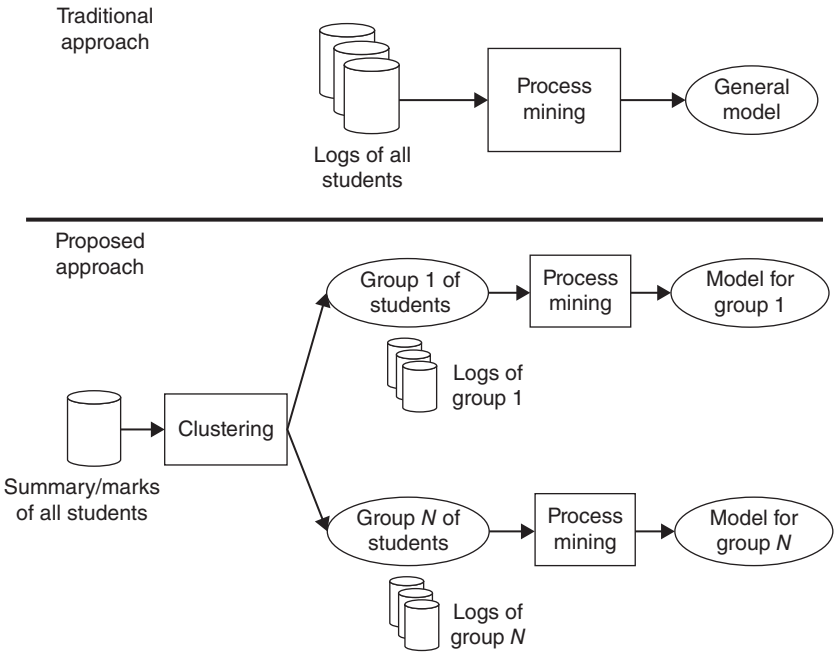


Figure 1.5 Representation of the proposed approach versus the traditional approach.

categorical value using Spain's traditional academic grading system: *fail* (from 0 to 4.9) and *pass* (from 5 to 10). By applying this manual clustering approach, two groups are easily detected from the 84 students:

- 16 students whose final marks were less than 5 (*fail*)
- 68 students whose final marks were greater than or equal to 5 (*pass*)

Automatic clustering uses the Moodle usage data, which were obtained after students worked on the course. We mainly used the reports or summaries of each student's interactions in Moodle. It is important to note that we have only used selected variables and that we have filtered the actions in our log file. The variables selected can be grouped into four different types (see Tables 1.4, 1.5, 1.6, and 1.7) by taking into account what they represent on a higher granularity level: variables related to time spent working, variables related to procrastination, variables related to participation in forums, and other variables.

This LMS version stores a total of 76 variables, but looking at previous works in the same line, only 12 actions make sense when representing the students' performance in the Moodle course used for the experiment. Some of these actions are extracted directly from Moodle records; however, to make sense of the data, it is sometimes advisable to formulate queries to obtain aggregated results. Therefore, other variables are calculated based on those records using a simple operation; for example, as seen in Table 1.5, the variable *days task* is calculated by subtracting the date on which the student uploaded the task (Moodle records this automatically from

TABLE 1.4 Variables related to the time spent working

Name	Description	Extraction Method and Moodle Nomenclature	Additional Information
Time theory	Total time spent on theoretical components of the content	Sum of the periods between resource view and the next different action	Students have a period of 15 days to learn from theory. The number of units is designed to be implemented during one semester on a weekly basis. Students have available one different unit every Monday. The theoretical contents remain available gradually till the end of the program
Time task	Total time spent in instructional tasks	Sum of the periods between *quiz view*/*quiz attempt*/*quiz continue attempt*/*quiz close attempt* and the next different action	Students have a period of 15 days to complete the tasks. The number of units is designed to be implemented during one semester on a weekly basis. A period of 15 days is imposed to do the task in order to coregulate them and avoiding procrastination
Time forums	Total time spent reviewing forums	Sum of the periods between *forum view* and the next different action	Students have a period of 15 days to go through their peers' comments and post at the forums

TABLE 1.5 Variables related to procrastination

Name	Description	Extraction Method and Moodle Nomenclature	Additional Information
Days theory	How many days in a 15-day period the students wait to check the content at least once (in days)	Date of *resource view* since the content is available	Students have a period of 15 days to learn from texts
Days tasks	How many days in a 15-day period the students wait to check the task at least once (in days)	Date of *task view* since the task is available	Students have a period of 15 days to complete the tasks
Days "hand in"	How many days in a 15-day period take the students to complete the task (in days)	Date of *quiz close attempt* since the task is available	Students have a period of 15 days to complete the tasks
Days forum	How many days in a 15-day period the students wait to check the forum (in days)	Date of *forum view forum* or *forum view discussion* since the forum is available	Students have a period of 15 days to go through their peers' comments
Days posting	How many days in a 15-day period take the students to post in the forum (in days)	Date of *forum add discussion* or *forum add replay* since the forum is available	Students have a period of 15 days to post at the forums

TABLE 1.6 Variables related to participation in forums

Name	Description	Extraction Method and Moodle Nomenclature	Additional Information
Words fórums	Number of words in forum posts	Extracting the number of *forum add discussion* and *forum add replay* words	There is no restriction in the number of words that students can use in the post
Sentences fórums	Number of sentences in forum posts	Extracting the number of *forum add discussion* and *forum add replay* sentences	There is no restriction in the number of sentences that students can use in the post

TABLE 1.7 Other variables

Name	Description	Extraction Method and Moodle Nomenclature	Additional Information
Relevant actions	Number of relevant actions in the LMS	Total of relevant actions considered	Actions like log in, log out, profile updating, check calendar, refresh content, etc. are dismissed
Activity days	Number of days between the first and the last relevant action on every unit, for example, first action *check quiz* and last action 5 days later *quiz close attempt*	Date of last relevant action—date of first relevant action (in days)	Students have a period of 15 days to complete the unit

the module *assignment* and the variables related to the *actions*), from the date it was possible to view the task (Moodle also records this automatically from the module *assignment* and the variables related to the *views*).

Other reasonable variables are easily extracted through similar procedures. For example, variables such as *time spent on theoretical contents* or *time spent in forums* are not as reliable as we would like because the experience took place outside of teaching hours; thus, while using Moodle, the students could simultaneously be working or surfing the Internet. However, the variable *time spent in tasks* is a reliable indicator for this course because the Moodle module quiz allows a time limit to be set for every task. Here, we see some of the added difficulties of being out of the laboratory and in a real educational condition. In this regard, the time variables by themselves are very tricky. It might seem that the more time those students spend studying, the better grades they should receive, but the relationship is not as simple as this; it mainly depends on the quality of the studying time. For that reason, the value of these variables is necessarily linked to other relevant variables in the learning progress, such as the groups used in the automatic clustering.

Other examples of feasible indicators of the students' performance are the variables *typing time* and *number of words in* forums; the latter was selected for this tutorial. Based on common sense, the variable *typing time* could be mediated by a student's individual skills. Nevertheless, according to the literature, variables such as

number of messages sent to the forum or *number of forum messages read* are related to student achievement. Accordingly, we think that the mean *number of words* and *sentences* in posts would be a good indicator of the quality of the answers because students are asked to post a reflection. On that basis, we have chosen what we think are the most representative and objective variables available in the Moodle logs.

Next, it is necessary to transform or convert all this information (from Tables 1.4, 1.5, 1.6, and 1.7) into an attribute-relation file format (ARFF) summary file. This is the data format used by Weka (Witten et al., 2011), which is the DM tool used for the study's clustering. Weka is a collection of machine learning algorithms for DM tasks. The Weka system has several clustering algorithms; we used the expectation–maximization (EM) clustering algorithm. This algorithm is used in statistics to find maximum likelihood estimators of parameters in probabilistic models that rely on unobservable variables. We have selected this specific algorithm because it is a well-known clustering algorithm that does not require the user to specify the number of clusters. Our objective is to group together students who have similar characteristics when using Moodle. In order to do this, we used Weka Explorer (see Fig. 1.6): in the "Preprocess" tab, we clicked on the "Open file…" button and selected the previous summary .ARFF file. Then, we clicked on the "Cluster" tab and, in the "Clusterer" box, selected the "Choose" button. In the pull-down menu, we selected the cluster scheme "EM" and then clicked on the "Start" button to execute the algorithm.

When the training set was complete, the "Cluster" output area on the right panel of the "Cluster" window was filled with text describing the results of the

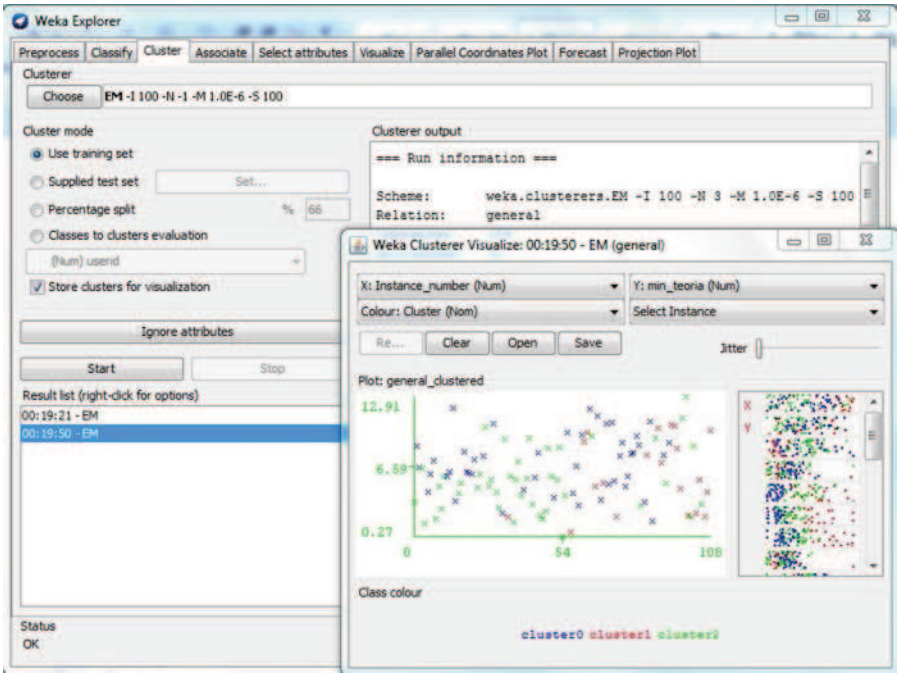


Figure 1.6 Weka clustering interface.

TABLE 1.8 Values (mean ± std.dev.) of the centroids of each cluster

Attribute	Cluster 0	Cluster 1	Cluster 2
Time theory	5.9 ± 3.2	7.1 ± 2.6	3.9 ± 1.5
Time tasks	14.1 ± 2.7	11.3 ± 6.2	5.3 ± 1.9
Time forums	8.7 ± 7.0	12.4 ± 5.4	7.6 ± 4.5
Days theory	6.0 ± 2.4	1.6 ± 6.9	8.4 ± 2.6
Days tasks	3.5 ± 1.0	1.8 ± 0.8	6.8 ± 2.3
Days "hand in"	4.8 ± 0.9	3.0 ± 1.2	9.3 ± 2.2
Number of words in forums	7.5 ± 6.5	9.2 ± 3.5	5.3 ± 3.4
Number of sentences in forums	92.9 ± 23.1	107.8 ± 40.6	78.1 ± 39.4

training and testing. In our case, we obtained three clusters, with the following distribution of students:

1. *Cluster 0*: 23 students (22 pass and 1 fail)
2. *Cluster 1*: 41 students (39 pass and 2 fail)
3. *Cluster 2*: 20 students (13 fail and 7 pass)

Clustering algorithms provide a highly interpretable result model by means of the values of each cluster centroid (Table 1.8). The centroid represents the most typical case/student or prototype in a cluster, and it does not necessarily describe any given case in that cluster.

As shown by the mean values in the different variables, students in Clusters 0 and 1 (in which most students passed) obtained higher values than those in Cluster 2 (in which most students failed) in terms of times (for theory, task, and forums) and counts (of words and sentences in forums). However, Clusters 0 and 1 had lower values for days (related to theory, tasks, and assignments hand in). Cluster 0 gives priority to the procedural level of knowledge, corresponding to the scores of the time and days tasks. The students comprising that cluster also seemed to show an achievement or strategic approach based on the prioritization of actions related to the compulsory assignments. In contrast, students belonging to Cluster 1 were presumably adopting a more dedicated approach to learning; note that the scores were good whether the variables were related to compulsory or suggested assignments. Finally, Cluster 2, comprised of students who normally fail, shows maladaptive variable levels and less frequent activity in the LMS.

1.3 WORKING WITH ProM

As is well known, instructors can easily gain insight into the way students work and learn in traditional learning settings. However, in LMS, it is more difficult for teachers to see how the students behave and learn in the system and to compare that system to other systems with structured interactions. These environments provide data on the interaction at a very low level. Because learner activities are crucial for an effective online teaching–learning process, it is necessary to search for empirical and effective tools to better observe patterns in the online environment; EPM and particularly ProM could be good resources for this purpose. Furthermore, the creation and evaluation of the models generated with ProM allow the researcher or instructor to not

only know more about the learning results but also to go through the learning process to better understand the referred results. Therefore, we used ProM (Van der Aalst, 2011a) for EPM. ProM is an extensible framework that supports a wide variety of PM techniques in the form of plug-ins.

Among the wide variety of algorithms, we applied the robust algorithm Heuristic Miner (Weijters, van der Aalst, & de Medeiros, 2006) to investigate the processes in the users' behavior. In this context, Heuristic Miner can be used to express the main behavior registered in an event log. It focuses on the control-flow perspective and generates a process model in the form of a Heuristics Net for the given event log. Therefore, the Heuristic Miner algorithm was designed to make use of a frequency-based metric that is less sensitive to noise and the incompleteness of the logs. As quality measures, we used fitness and the default threshold parameters of the Heuristic Miner algorithm. We applied Heuristic Miner using the ProM tool over the six previously obtained log data sets in order to discover students' process models and workflows. We applied the algorithm to each of these logs:

1. All students (84 students)
2. Students who passed (68 students)
3. Students who failed (16 students)
4. Students assigned to Cluster 0 (22 pass and 1 fail)
5. Students assigned to Cluster 1 (39 pass and 2 fail)
6. Students assigned to Cluster 2 (13 fail and 7 pass)

The first task after starting ProM was to import a log file in the following way: Click the "import…" icon in the upper-right corner and select the appropriate MXML file. The result is shown in Figure 1.7.

Figure 1.7 ProM interface for importing a log file.

Next, we could apply all kinds of ProM plug-ins. We could access all the available plug-ins by clicking in the ► tab in the upper middle bar (see Fig. 1.8).

From the list of plug-ins available in ProM, we selected "Mine for a Heuristic Net using Heuristics Miner" and click the "Start" button. Then, the parameters of Heuristic Miner were shown (see Fig. 1.9). The default values of these parameters were used in all our experiments.

Figure 1.8 List of plug-ins available in ProM.

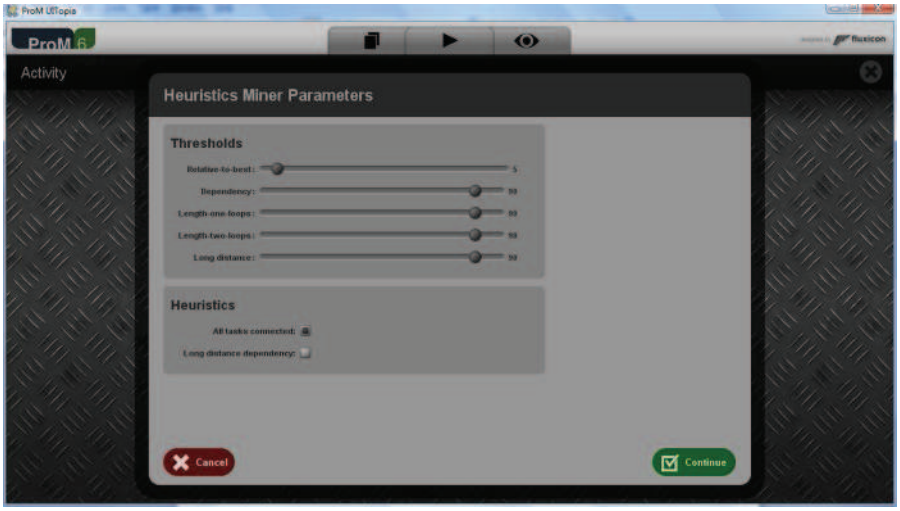


Figure 1.9 Parameters of the Heuristics Miner.

Once we pressed the "Continue" button on the configuration parameters screen, the discovered models are shown.

1.3.1 Discovered Models

The model discovered by the Heuristics Miner algorithm (Van der Aalst, 2011b) is a heuristic network that is a cyclic, directed graph representing the most common behaviors of students browsing the course. In this graph, the square boxes represent the actions of the students when interacting with Moodle's interface, and the arcs/links represent dependences/relations between actions. Next, we describe the discovered models using each of our log files.

Figure 1.10 shows the heuristic network obtained when using the log file with all students. We can see that there are two subnets that most of the students follow in the course. The upper subnet consists of *view forum* actions about the most viewed forums in the course, and the lower subnet consists of *view quiz* actions about the most viewed quizzes in the course. From the expert point of view, this information, although useful, is only a surface-level approach to the learning process, which we want to explore more deeply. It is important to note that these networks show the general behavior of all the students (fail and pass students mixed); probably for this reason, there are a lot of relations/dependences between the actions that make the model harder to interpret.

Figure 1.11 shows the heuristic network obtained when using only the logs of the *passing students*. We can see that this type of student followed a relatively high number of subnets. Respectively, these students followed seven subnets: *quiz view*, *quiz view summary*, *quiz attempt*, *quiz close attempt*, *quiz continue attempt*, *quiz review*, and *forum view*. Thus, we can see that passing students were very active in quiz actions. This makes sense, as based on the instructor's directions, success in the course was oriented to practical tasks. This model provides better insight into the students' learning processes than the previous model did. The students who passed the course performed well in the core subprocesses of learning: collaborative learning (*forum view*), forethought (*quiz view, quiz view summary*), performance (*quiz attempt*, *quiz close attempt, quiz continue attempt*), and self-reflection (*quiz review*).

Figure 1.12 shows the heuristic network obtained when using only the logs of *failing students*. This figure shows the two subnets used by most of the students who failed the course. The top subnet consists of *page view* actions for the most viewed pages in the course's text content. The bottom subnet consists of *view quiz* actions for the most viewed quizzes. Taking into account that the actions related to practical tasks and forums are not especially notable, we could conclude that, rather than engaging in the task and talk in the forums, the failing students could be using their time to study the theoretical contents. However, as previously mentioned, the course's goal was not the acquisition of declarative knowledge but putting this knowledge into practice. Based on this simple fact, these students' learning is incorrectly oriented.

It is also interesting to see how the heuristic net of students who failed is much smaller than those of the heuristic net for all students and for passing students. On the

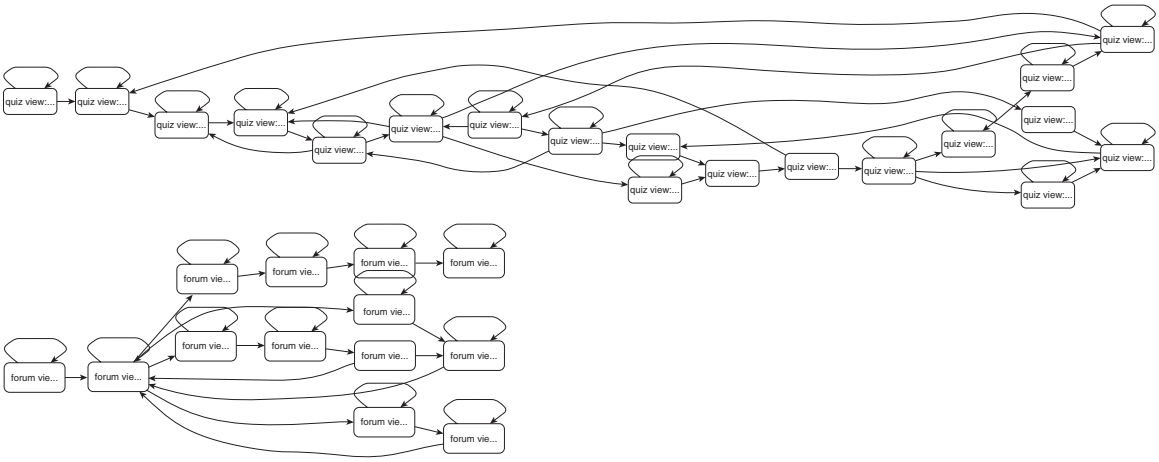


Figure 1.10 Heuristic net of all students.

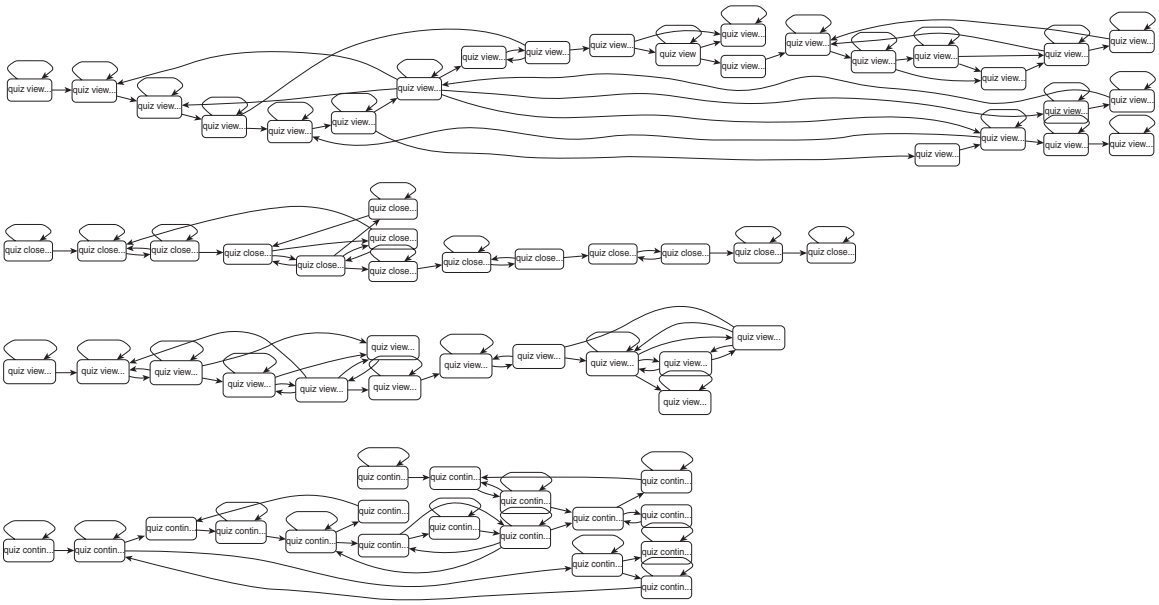


Figure 1.11 Heuristic net of passing students.

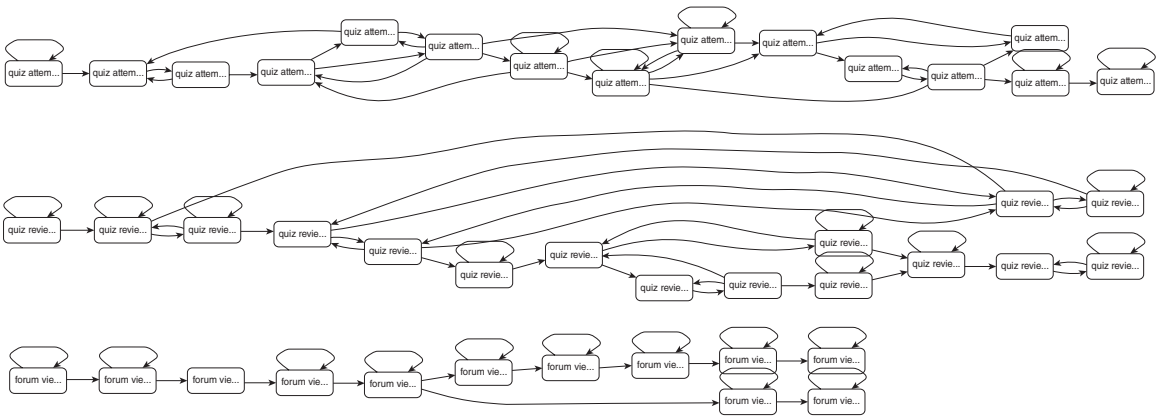


Figure 1.11 (*Continued*)

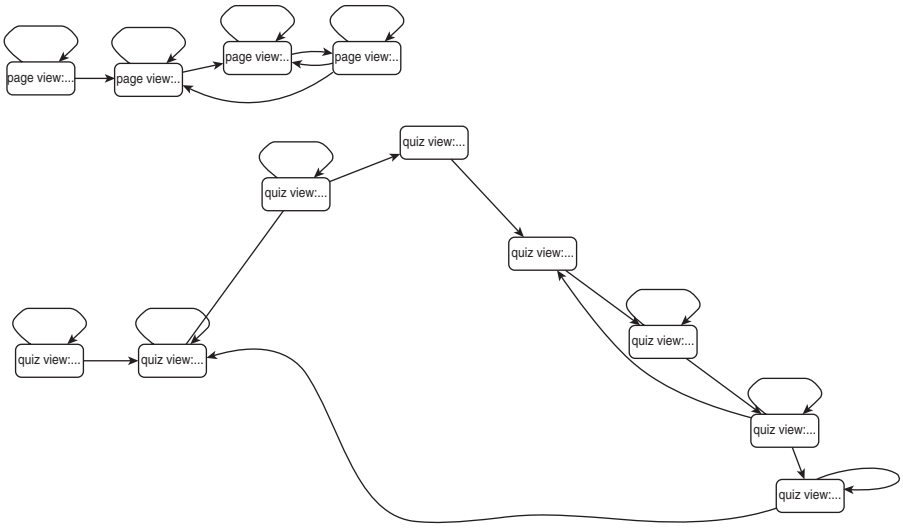


Figure 1.12 Heuristic net of failing students.

other hand, this chapter has already offered warnings about wrong assumptions related to study time. This could also be applied to the number of actions; having more interactions with the LMS does not necessarily reflect better performance. However, the evident scarcity of interaction with the learning environment showed by failing students in this particular case is conclusive. In the end, based on this chapter's scope, there is good news, as the behavior of failing students would be easy to detect and interpret.

Finally, we show one example of the three clusters. Figure 1.13 shows the heuristic net obtained when using only the logs of passing students in Cluster 0. This figure shows the five subnets followed by this subtype of students. According to this model of passing students from Cluster 0, the obtained subnets consist of these actions: *quiz view*, *quiz close attempt*, *quiz review*, *quiz continue attempt*, and *forum view*. If we compare Figure 1.11 (all students who pass) with Figure 1.13 (a subtype of students who pass), we can see that not only are there fewer subnets (five instead of seven) but also that the networks are smaller (with a smaller number of nodes and arcs). The usefulness of this, apart from generating models that are easier to interpret, is that the instructor of the course knows and can select which are the crucial variables for successful performance in the course and/or which target variables will determine the ProM model generation from the previous clustering.

1.3.2 Analysis of the Models' Performance

We have also carried out an analysis of the performance of the previously obtained models (heuristic networks). In order to do this, a fitness measure is normally used (Ayutaya et al., 2012). Fitness is a quality measure indicating the gap between the

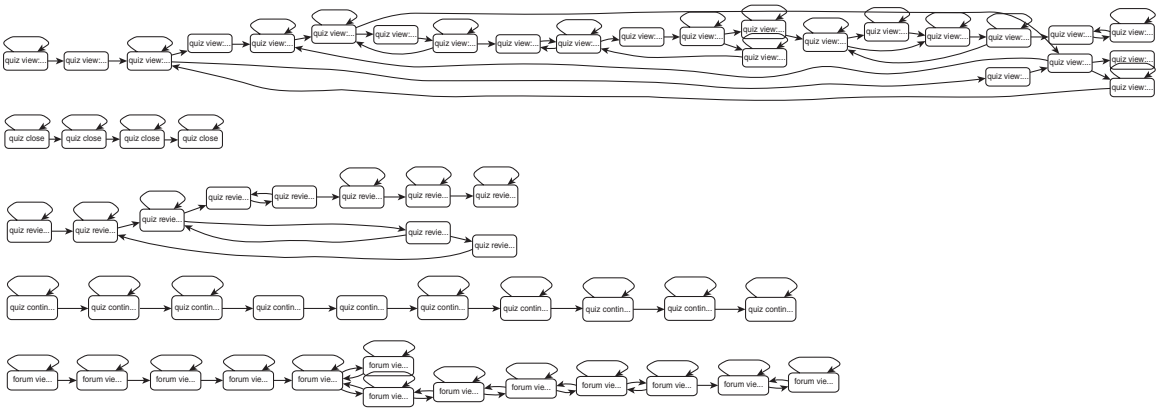


Figure 1.13 Heuristic net of Cluster 0 students.

behavior actually observed in the log and the behavior described by the process model. It gives the extent to which the log traces can be associated with the execution paths specified by the process model. If a model has a poor fitness value, this indicates that the mined process model does not successfully parse most of the log traces. The fitness results that we have obtained for each of our models are shown in Table 1.9.

As we can see in Table 1.9, the lowest fitness was obtained when using all data, for which 70 of 84 students fit to the obtained model (83.33%). On the other hand, all the other models (using both manual and automatic clustering) obtained a fitness value greater than 90% in all cases. The highest fitness value was obtained when using data from students who failed, for which 15 of 16 students fit the obtained model (93.75%). Thus, in this case, we can see that these specific models, which were obtained using manual and automatic grouping/clustering, performed better than the general model obtained from all students.

Additionally, we evaluated the compressibility of the obtained models. In order to do this, the complexity or size of each model is normally used (Bogarín et al., 2014). We have used two typical measures from graph theory: the total number of nodes and the total number of links in the models/graphs (see Table 1.10).

As we can see in Table 1.10, the smallest (most comprehensible) model was obtained with the data set of failing students, followed by those for all students and Cluster 2 students. The other three obtained models were much larger and more complex, especially the model using passing students.

TABLE 1.9 Fitness of the obtained models

Data Set	Fitness
All students	0.8333
Pass students	0.9117
Fail students	0.9375
Cluster 0 students	0.9130
Cluster 1 students	0.9024
Cluster 2 students	0.9000

TABLE 1.10 Complexity/size of the obtained models

Data Set	Number of Nodes	Number of Links
All students	32	70
Pass students	113	244
Fail students	12	24
Cluster 0 students	61	121
Cluster 1 students	59	110
Cluster 2 students	38	84

1.4 CONCLUSION

The present work proposed using clustering to improve EPM and, at the same time, optimize both the performance/fitness and comprehensibility/size of the model. We have obtained different models by using data sets from different groups of students:

- In the model from the data set of all students, the students showed different behavior and only had a few common actions because there was a mix of different types of students (*passing* and *failing*).
- In the model from the data set of students who failed in Cluster 2 students, the students only showed a few common behavioral patterns because these types of students (who failed) were less participatory or interactive when browsing the course than were the others (who passed).
- In the model from the data set for students who passed in Cluster 0 and Cluster 1, the students showed a much higher number of common behavioral patterns because these types of students (who passed) were more active users than the others (who failed).

From an educational and practical point of view—to be able to use this information for providing feedback to instructors about student learning—these models could easily be used to point out which new students are at risk of failing a course. For example, instructors only have to check to see which new students are following the same routes/behavioral patterns shown by the heuristic net of students who failed.

On the other hand, model comprehensibility is a core goal in education due to the transfer of knowledge that this entails. Making graphs, models, and visual representations understandable to teachers and students makes these results essential for monitoring the learning process and providing feedback; one of our goals is to do precisely that in real time. Furthermore, Moodle does not provide specific visualization tools for students' usage data that would allow the different agents involved in the learning process to understand the large amount of raw data and to become aware of what is happening in distance learning. In addition the results can also be extended to the improvement of adaptive hypermedia learning environments, for which prompting the students about recommended learning paths, shortcuts, etc. is the basis for enhancing the learning experience in a more strategic way.

Finally, in the near future, more experiments will be conducted to test our approach using other types of courses from different fields of knowledge. We also want to explore other ways to group students before PM. For example, we could group students based on the triangulation of different sources of information, such as self-reported data or psychophysiological measures based on students' metacognitive behavior. We also want to test if clustering the content (or even a manual semantic mapping of the content/course structure) would allow us to simplify the process models. Even further, we could split the course into semantic blocks and see how students progress—either as a logical progression (e.g., unit 1.8) or as a time sequence (e.g., week 1–10). This would be a very interesting way to identify faults in the process. For example, students progressing through units in similar ways are more likely to perform better (i.e., they have a strategy).

ACKNOWLEDGMENTS

This research is supported by projects of the Spanish Ministry of Science and Technology TIN2014-55252-P, EDU2014-57571-P, and the European Funds for Regional Development Ref. GRUPIN14-053.

REFERENCES

Ayutaya, N. S. N., Palungsuntikul, P., & Premchaiswadi, W. Heuristic mining: Adaptive process simplification in education. International Conference on ICT and Knowledge Engineering, 221–227, November 21–23, 2012.

Azevedo, R., Behnagh, R., Duffy, M., Harley, J. M., & Trevors G. J. Metacognition and self-regulated learning in student-centered learning environments. In Jonassen, D. H. & Land, S. M. (eds), *Theoretical Foundations of Student-Center Learning Environments*, 2nd edition. Erlbaum, Mahwah, NJ, 216–260, 2012.

Bannert, M., Reimann, P., & Sonnenberg, C. Process mining techniques for analysing patterns and strategies in students' self-regulated learning. *Metacognition and Learning*, 9(2), 161–185, 2014.

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. Clustering for improving educational process mining. International Conference on Learning Analytics and Knowledge, New York, 11–15, 2014.

Boiarsky, C. A model for analyzing revision. *Journal of Advanced Composition*, 5, 65–78, 1984.

Cairns, A. H., Ondo, J. A., Gueni, B., Fhima, M., Schwarcfeld, M., Joubert, C., & Khelifa, N. Using semantic lifting for improving educational process models discovery and analysis. *CEUR Workshop Proceedings*, 1293, 150–161, 2004.

Cairns, A. H., Gueni, B., Fhima, M., Cairns, A., David, S., & Khelifa, N. Towards custom-designed professional training contents and curriculums through educational process mining. IMMM 2014, The Fourth International Conference on Advances in Information Mining and Management, 53–58, 2014.

Dutt, A., Aghabozrgi, S., Ismail, M. A. B., & Mahroeian, H. Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, 5, 112–116, 2015.

Romero, C., Ventura, S., & García, E. Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), 368–384, 2008.

Schoor, C. & Bannert, M. Exploring regulatory processes during a computer-supported collaborative learning task using process mining. *Computers in Human Behavior*, 28(4), 1321–1331, 2012.

Sedrakyan, G., Snoeck, M., & De Weerdt, J. Process mining analysis of conceptual modeling behavior of novices–empirical study using JMermaid modeling and experimental logging environment. *Computers in Human Behavior*, 41, 486–503, 2014.

Southavilay, V., Yacef, K., & Calvo, R. A. Process mining to support student's collaborative writing. Educational Data Mining Conference, 257–266, 2010.

Trcka, N. & Pechenizkiy, M. From local patterns to global models: Towards domain driven educational process mining. International Conference on Intelligent Systems Design and Applications, Milan, Italy, 1114–1119, 2009.

Van der Aalst, W. M. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, Berlin/Heidelberg/New York, 2011a.

Van der Aalst, W. M. Using process mining to bridge the gap between BI and BPM. *IEEE Computer*, 44(12), 77–80, 2011b.

Van der Aalst, W. M., Adriansyah, A., de Medeiros, A. A., Arcieri, F., Baier, T., Blickle, T., & Pontieri, L. Process mining manifesto. Business Process Management Workshops, 169–194, 2012.

Vellido, A., Castro, F., & Nebot, A. Clustering educational data. In Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. (eds), *Handbook of Educational Data Mining*. CRC Press, Boca Raton, FL, 75–92, 2011.

Weijters, A. J. M. M., van der Aalst, W. M., & de Medeiros, A. A. Process mining with the Heuristics Miner-algorithm. Technische Universiteit Eindhoven, Tech. Rep. WP, 166, 2006.

Witten, I. H., Frank, E., & Hall, M. A. *Data Mining, Practical Machine Learning Tools and Techniques*, 3rd edition. Morgan Kaufman Publishers, Amsterdam, 2011.

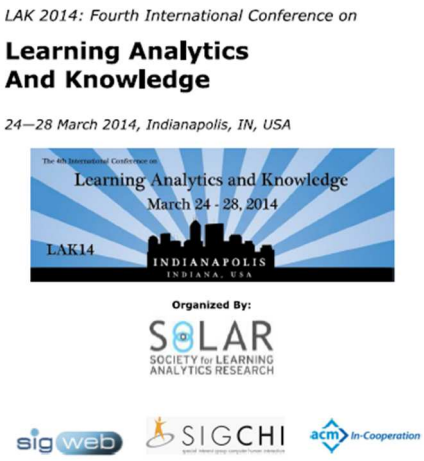
ARTÍCULO 4

Referencia:

- Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In: *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge.* Indianápolis, USA. 11-15.

Medidas de Calidad Científica:

- CORE B
- Learning Analytics And Knowledge
- Número total de citas en GoogleShoolar: 45



Clustering for improving Educational Process Mining

Alejandro Bogarín, Cristóbal Romero
Department of Computer Science
University of Cordoba, Spain
i02bovea@uco.es, cromero@uco.es

Rebeca Cerezo, Miguel Sánchez-Santillán
Department of Psychology
University of Oviedo, Spain
cerezorebeca@uniovi.es,
melsanchezsantillan@gmail.com

ABSTRACT
In this paper, we propose to use clustering to improve educational process mining. We want to improve both the performance and comprehensibility of the models obtained. We have used data from 84 undergraduate students who followed an online course using Moodle 2.0. We propose to group students firstly starting from data about Moodle's usage summary and/or the students' final marks in the course. Then, we propose to use data from Moodle's logs about each cluster/group of students separately in order to be able to obtain more specific and accurate models of students' behaviour. The results show that the fitness of the specific models is greater than the general model obtained using all the data, and the comprehensibility of the models can be also improved in some cases.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data Mining; K.3.1 [Computer Uses in Education]: Computer-assisted instruction (CAI), Computer-managed instruction (CMI);

General Terms
Algorithms, performance, experimentation.

Keywords
Clustering, process mining, educational data mining, learning analytics.

1. INTRODUCTION
Educational Data Mining (EDM) and Learning Analytics and Knowledge (LAK) [10] study data and analytics in education, teaching, and learning, suggesting educational priorities and undertaking high-quality research into the models, methods, technologies, and impact of analytics. One of the current promising techniques in EDM and LAK is Educational Process Mining (EPM). The basic idea of process mining is to extract knowledge from event logs recorded by an information system. EPM [12] aims to (i) construct complete and compact educational process models that are able to reproduce all observed behaviour, (ii) to check whether the modeled behaviour matches the observed behaviour, and (iii) to project information extracted from the logs onto the model, to make unexpressed knowledge explicit and to facilitate better understanding of the process.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA

http://dx.doi.org/10.1145/2567574.2567604

The results of EPM can be used to get a better understanding of the underlying educational processes, to generate recommendations and advice to students, to provide feedback to either students, teachers or/and researchers, to detect learning difficulties early, to help students with specific learning disabilities, to improve management of learning objects, etc.; but crucially, helping solve the difficulties that students of different ages show when learn in highly cognitively and metacognitively demanding learning environments like hypermedia or Computer Based Learning Environments [3]. However, we have found two problems when using EPM: 1) the model obtained cannot fit well to the general students' behaviour and 2) the model obtained can be too large and complex for use or analysis by an instructor. In order to resolve these problems, we propose to use clustering to improve both the fitness and comprehensibility of the obtained models by EPM.

This paper is organized as follow. Section 2 is a background about related works. Section 3 describes the proposed approach. Section 4 describes the datasets used. Section 5 describes the experiments and finally, section 6 shows the conclusion and future works.

2. BACKGROUND
Most of the traditional data mining techniques focus on data dependencies or simple patterns and do not focus on the process as a whole and do not provide visual representation of the complete educational process ready to be analyzed [6]. However, EPM techniques aim to extract process-related knowledge from event logs recorded by an information system [13]. In fact, nowadays there are an increasing number of examples about the applicability of process mining in education.

Process mining has been used to extract knowledge from a particular type of an educational information system, considering (oversimplified) educational processes reflecting student behaviour only in terms of their examination scores [13]. Several process mining techniques such as Petri nets, Heuristic and Fuzzy miner have been used to analyse assessment data from recently organized online multiple choice tests [6]. Petri nets analysis has been used to find learning paths or to optimise the path that a student must follow in order to reach a degree or a qualification [4]. Bottleneck mining and petri net simulation have been used to detect discrepancies between the flows prescribed in a student's registration model and the actual process instances [2]. Heuristic Mining has been used to analyze students' writing activities to improve not only the quality of the written document but more importantly the writing skills of those involved [11]. Heuristic Mining has been also used to investigate the processes in students' registration of Thailand's universities for adaptive process simplification in education [1]. Heuristic mining has been

also used to analyse learning paths in LMS and track learning behaviour in relation to respective learning styles, which must be identified in advance. Finally, we want to notice that we haven't found any work that uses clustering techniques together with EPM.

3. PROPOSED APPROACH

In this paper, we proposed an approach that uses clustering for improving educational process mining (see Figure 1).

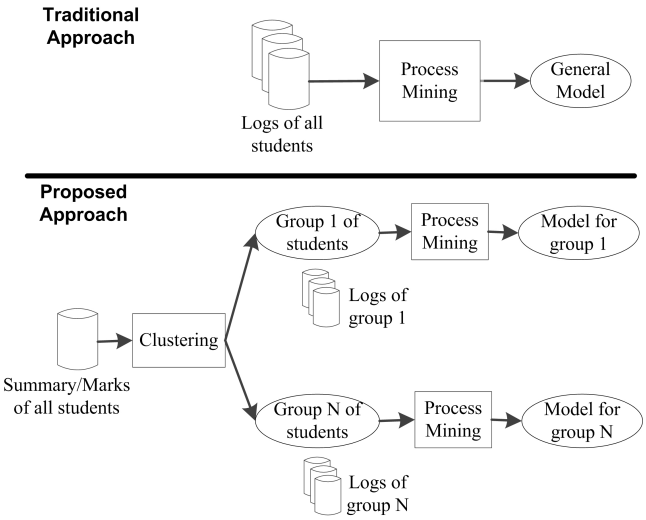


Figure 1. Proposed approach versus the traditional approach.

The traditional approach uses all event log data to reveal a process model of student's behaviour.

On the other hand, this proposed approach firstly applies clustering in order to group students with similar marks or characteristics. It then applies process mining to discover more specific models of student behaviour. We propose two different clustering/grouping approaches:

- **Manual**: To group students directly using only the students' marks obtained in the final exam of the course.
- **Automatic**: To group students using a clustering algorithm over the student's interaction with the Moodle's course.

4. DESCRIPTION OF THE DATA USED

The datasets used in this research were gathered from a Moodle 2.0 course used by 84 undergraduate university students taking the Psychology degree in a university in the north of Spain. The study was conducted during two different semesters in the years 2011-2012 and 2012-2013. The experiment took the form of an assignment in the curriculum of a 3rd year compulsory subject. Students were asked to participate in an eLearning/training programme about "learning to learn", related to the subject's topic, completed entirely out of teaching hours. The programme was made up of 11 different units that were sent to the students on a weekly basis, but each of them was able to work on it during a 15 day period. Each unit was composed of three different types of contents:

- **Declarative knowledge level**: Theoretical contents, description, information, and how-to put the "learn to learn" strategy or strategies of the week into practice.
- **Procedural knowledge level**: Practical tasks where the students had to put their declarative knowledge into practice.
- **Conditional knowledge level**: Discussion forums where the students had to discuss about how they had or would use the strategy or strategies of the week in different contexts.

Students get an extra point in their final subject grade if they complete at least 80% of the assignments. Compulsory assignments for each unit were to solve the practical task and to post at least one comment on each unit forum. Suggested assignments for each unit were to understand the theoretical contents, put them in practice in the task, and share their experience about the week's topic in the forum. Based on the data obtained after students worked on the entire programme, we used three different sources of information.

On the one hand, we used a summary about the interaction of each student in Moodle calculated from Moodle's log and different tables of the database (Table 1).

Table 1. Variables of a student in our Moodle summary file.

Name	Description	Extraction Method under Moodle nomenclature
Time Theory	Total time spent on theoretical contents	Time between viewing resource and next different action
Time Tasks	Total time spent on practical tasks	Time between viewing quiz/ attempting quiz/ continuing quiz attempt/ closing quiz attempt and next different action
Time Forums	Total time spent reviewing forums	Time between viewing forum and the next different action
Days Theory	How long students wait to check the content	Date resource was viewed since content became available on Moodle (in days)
Days Tasks	How long students wait to check the task	Date task was viewed since task became available on Moodle (in days)
Days "hand in"	Taking time in hand in the task	Date quiz attempt closed since task became available on Moodle (in days)
Number of Words in forums	Number of words in forum posts	Extracting number of words added to forum discussion OR forum replay words added
Number of sentences in forums	Number of sentences in forum posts	Extracting number of sentences added to forum discussion OR forum replay

		sentences added

We saved/converted all this information into an .ARFF (Attribute-Relation File Format) file to be able to apply clustering algorithms provided by Weka [16].

On the other hand, we also used the log file provided by Moodle directly (see Table 2).

Table 2. Variables of a Modle log files.

Attribute	Description
Course	The name of the course
IP Address	The IP of the device used to access
Time	The date they accessed it
Full Name	The name of the student
Action	The action that student has done
Information	More information about the action

We preprocessed this log file (see Table 2) thus: we only used the last four variables, that is, we did not use the name of the course (it is the same for all records) or the IP address (it is irrelevant for our purposes). We turned the students' names into IDs (Identifiers) to maintain their privacy. We filtered the possible actions in our log file. So, from 39 actions that Moodle stored in our log file, we only used the next 20 actions that are related to the students activities in the course: upload assignment, view assignment, view course, view folder, add forum discussion, add forum post, update forum post, view forum discussion, view forum, view page, submit questionnaire, view questionnaire, attempt quiz, close quiz attempt, continue quiz attempt, review quiz, view quiz, view quiz summary, view resource, and view url. We then transformed this log file into a MXML (Minimal XML) file using the proMimport framework in order to can use ProM [14] for later process mining.

Finally, we used the students' final marks, or the scores they obtained in the final exam at the end of the course. This is a file provided by the instructor that contains both each student's ID and his or her final mark (a numeric value on a 10-point scale). We turned this continuous value into a categorical value using the traditional academic grading in Spain: 0-4.9: fail and 5-10: pass.

5. EXPERIMENTS

We carried out several experiments to test our proposal. In the first, we used all the log data about the 84 students. In the second, we divided firstly the original log file into two datasets: one that contains the 68 students who passed and other with the 16 students who failed. And in the last experiment, we have used the Expectation-Maximization (EM) clustering algorithm provided by Weka [16] to group together students of similar characteristics when using Moodle (see Table 1). We used this algorithm because it is a well-known clustering algorithm that does not require the user to specify the number of clusters to find. In our case, we obtained three clusters with the following distribution of students:

- **Cluster 0**: 23 students (22 pass and 1 fail).
- **Cluster 1:** 41 students (39 pass and 2 fail).
- **Cluster 2**: 20 students (13 fail and 7 pass).

Clustering algorithms provide a high interpretable result model by means of the values of each cluster centroid (Table 3). The centroid represents the most typical case/student or prototype in a cluster, which does not necessarily describe any given case in that cluster.

Table 3. Values (mean±std.dev.) of centroids of each cluster.

Attribute	Cluster 0	Cluster 1	Cluster 2
Time Theory	5.9±3.2	7.1±2.6	3.9±1.5
Time Tasks	14.1±2.7	11.3±6.2	5.3±1.9
Time Forums	8.7±7.0	12.4±5.4	7.6±4.5
Days Theory	6.0±2.4	1.6±6.9	8.4±2.6
Days Tasks	3.5±1.0	1.8±0.8	6.8±2.3
Days "hand in"	4.8±0.9	3.0±1.2	9.3±2.2
Number of words in forums	7.5±6.5	9.2±3.5	5.3±3.4
Number of sentences in forums	92.9±23.1	107.8±40.6	78.1±39.4

As shown by the mean values in the different variables, students in clusters 0 and 1 (in which most students pass) obtain higher values than cluster 2 (in which most students fail) in times (theory, task and forums) and numbers (of words and sentences in forums), but lower values in days (theory, tasks and hand in). Cluster 0 gives priority to the procedural level of knowledge, corresponding to the scores at the *Time tasks* and *Days Tasks.* The students comprising that cluster also seem to show an achievement or strategic approach based on the prioritisation of the actions related to the *compulsory assignments*. In contrast, students belonging to Cluster 1 are presumably adopting a more dedicated approach to learning: note that the scores are good whether or not the variables are related with *compulsory or suggested* assignments. Finally, Cluster 2, comprised of students who normally fail, shows maladaptive profiles and a more infrequent use of Moodle.

We then applied process mining over the different datasets (all, pass/mark and clusters) in order to discover students' process models and workflows in each one of these logs. We used ProM, [14] a generic tool for implementing process mining tools in a standard environment. In fact, we applied Heuristic Miner, [1] one of the robust algorithms to investigate the processes in users' behaviour. Heuristic Miner can be used to express the main behaviour registered in an event log. It focuses on the control flow perspective and generates a process model in the form of a Heuristics Net for the given event log. Therefore, the Heuristic Miner algorithm was designed to make use of a frequency based metric and so it is less sensitive to noise and the incompleteness of logs. We used the default threshold parameters of Heuristic Miner algorithm provided by ProM and the fitness as a quality measure (see Table 4). Fitness is a quality measure indicating the gap between the behaviour actually observed in the log and the

behaviour described by the process model. It gives the extent to which the log traces can be associated with execution paths specified by the process model. If model has a poor fitness value, this indicates that the mined process model does not successfully parse most of the log traces. This may be due to the presence of noise, resulting in dangling activities and missing connections. It is also possible that the parameter settings do not notice all connections.

Table 4. Fitness of the obtained models

Dataset	Fitness
All students	0.8333
Pass students	0.9117
Fail students	0.9375
Cluster 0 students	0.9130
Cluster 1 students	0.9024
Cluster 2 students	0.9000

As we can see in Table 4, the lowest fitness was obtained when using all data in which 70 of 84 students fit to the obtained model, that is, 83.33% of all students. On the other hand, all the other models (obtained using both manual and automatic clustering) obtained a fitness value greater than 90% in all the cases. And the highest fitness value was obtained when using data from students who failed, where 15 of 16 students fit to the obtained model, that is, 93.75% of students who failed. So, in this case/experiment we can see that these specific models obtained using manual and automatic grouping/clustering performed/fitted better than the general model obtained from all students.

Next, some information about the level of complexity or size of each one of the obtained models (Table 5) and two examples of obtained models (Figure 2 and 3) are described. The model discovered by Heuristic Miner algorithm is a heuristic net, drawn as a directed cycle graph, which represents the most frequent behaviours of the students of the dataset used.

We have used two typical measures from graph theory (the total number of nodes and the total number of links) in order to see the level of complexity of the models.

Table 5. Complexity/Size of the obtained models

Dataset	N.Nodes	N.Links
All students	32	70
Pass students	113	244
Fail students	12	24
Cluster 0 students	61	121
Cluster 1 students	59	110
Cluster 2 students	38	84

As we can see in Table 5, the smaller or more comprehensible model was obtained with students who failed followed by all students and cluster 2 students. On the other hand, the other three models are much bigger and complex. We think that the reasons for this may be that:

- In the dataset "all students", the students show different behaviour and only have some common actions because there are mixed different type of students (pass and fail students).
- In the datasets "students who failed and cluster 2 students", the students show only some common behavioural patterns because this type of student participates/interacts little with Moodle.
- In the datasets "students who pass, cluster 0 and cluster 1", students show much more common behavioural patterns because these types of students are more active users of Moodle

Finally, two examples of obtained models (heuristic nets) are described. The first example shows the heuristic net obtained when using all students (Figure 2) and the second when using fail students (Figure 3). In our heuristic nets the square boxes represent the actions of the students when interacting with Moodle's interface, and the arcs/links represent dependences/relations between actions

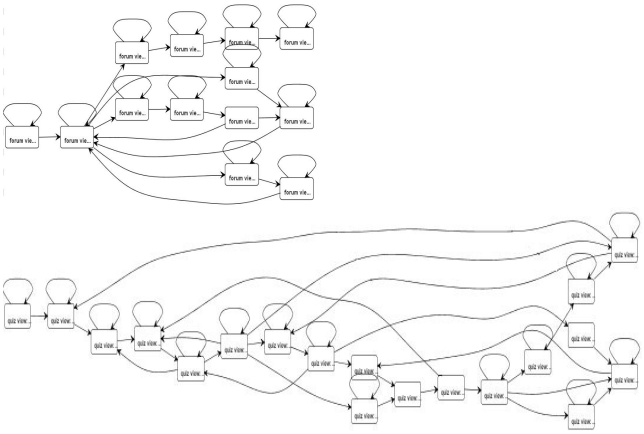


Figure 2. Heuristic net of all students.

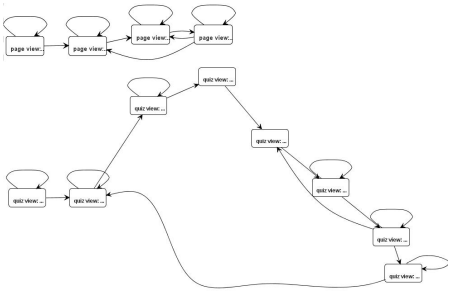


Figure 3. Heuristic net of fail students.

On the one hand, Figure 2 shows two subnets that follow most of the student of the course. The upper subnet consists of some view forum actions about the most viewed forums in the course. And the lower subnet consists of some view quiz actions about the most viewed quizzed in the course.

On the other hand, Figure 3 shows two subnets that follow most of the students who fail the course. The upper subnet consists of some page view actions about the most viewed pages. These pages are about general information on the course. The lower subnet consists of some view quiz actions about the most viewed

quizzed. In this case/dataset, the number of quizzes is much lower than in the previous case (see lower part of Figure 2) and not in the same exact order of visiting.

In summary, both Figures 2 and 3, show the most typical behaviour of the students in each case/dataset. But, it is interesting to see that the heuristic net of students who failed is smaller than the heuristic net of all students. From an educational and practical point of view (to be able to use this information for providing feedback to instructors about student learning), it could easily be used to point out new students at risk of failing the course. For example, instructors only have to check if new students follow the same specific routes/behavioural patterns that shown by the heuristic net of students who failed. That is, if they visit the same pages, view the same quizzes, and in the same order as previous students who failed.

6. CONCLUSIONS

In the present work, we propose to use clustering to improve educational process mining and, at the same time, optimise both the performance/fitness and comprehensibility/size of the model obtained. In particular, the comprehensibility of the model is a core goal in education due to the transferral of basic knowledge that it entails. Making graphs, models or visual representation more accessible or at least, accessible, to teachers and students, makes these results very useful for monitoring the learning process and providing feedback, one of our future goals being to do it in real time. Furthermore, Moodle does not provide specific visualization tools of students' usage data that let the different agents of the learning process understand these large amounts of raw data and become aware of what is happening in distance learning, apart from extending the use of the results to Adaptive Hypermedia Learning Environments in which it is very useful to prompt students or recommend learning paths, shortenings, etc., in order to enhance the learning experience in a more strategic way.

In the future, we want to do more experiments in order to test our proposed approach with other types of courses from different fields. We also want to explore other ways to group students before process mining. For example, grouping students based on the triangulation of the different sources of information used at this work and adding, if possible, self-report data from students' metacognitive behaviour.

7. ACKNOWLEDGMENTS

This work was supported by the Regional Government of Andalusia and the Spanish Ministry of Science and Technology projects, P08-TIC-3720, TIN-2011-22408 and EDU2010-16231, and FEDER funds.

8. REFERENCES

[1] Ayutaya, N. S. N., Palungsuntikul, P., & Premchaiswadi, W. 2012. Heuristic mining: Adaptive process simplification in education. *International Conference on ICT and Knowledge Engineering*. 221-227.

[2] Anuwatvisit, S., Tunggkasthan, A., Premchaiswadi, W. 2012. Bootleneck mining and petri net simulation in education situations. *Conference on ICT and Knowledge Engineering.* 244-251.

[3] Azevedo, R., Behnagh, R., Duffy, M., Harley, J. M., & Trevors G. J. 2012. Metacognition and self-regulated learning in student-centered learning environments. *Theoretical foun- dations of student-center learning environments*. Erlbaum, Mahwah, NJ, 2nd edition, 216–260.

[4] Campos-Rebelo, R., Costa, A., Gomes, L. 2012. Finding learning paths using petri nets modeling applicable to e-learning platforms. *International Federation for Information Processing*, 151-160.

[5] Holzhüter, M., Frosch-Wilke, D., Klein, U. 2013. Exploiting learner models using data mining for e-learning: a rule base approach. *Intelligent and Adaptive ELS*. Springer. 77-105.

[6] Pechenizkiy, M., Trcka, N., Vasilyeva, E., van der Aalst, W. M., & De Bra, P. 2009. Process Mining Online Assessment Data. *Educational Dada Mining Conference*, Cordoba, Spain, 279-288.

[7] Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaiane, O. 2009. Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE Transactions on Knowledge and Data Engineering*. 21(6): 759-772.

[8] Romero, C., Ventura, S. Zafra, A. and De Bra, P. 2009. Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computer&Education*, 53, 828-840.

[9] Romero, C., Lopez, M.I., Luna, J.M., and Ventura, S. 2013. Predicting students' final performance from participation in online discussion forums. *Computers&Education*. 68,458–472.

[10] Siemens, G., Baker, R.S.J.d.. 2012. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. *International Conference on Learning Analytics and Knowledge*. 1-3.

[11] Southavilay, V., Yacef, K., Calvo, R.A. 2010. Process mining to support student's collaborative writing. *Educational Data Mining Conference*, 257-266.

[12] Trcka, N., Pechenizkiy, M. 2009. From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining. *International Conference on Intelligent Systems Design and Applications*, Milan, Italy, 1114-1119.

[13] Trcka., N. Pechenizkiy, M. van der Aalst, W. 2011. Process mining from educational data. *Educational Data Mining Handbook*. CRC Press.

[14] Van der Aalst, W. 2011. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer.

[15] Weijters, A. J. M. M., van der Aalst, W. M., & De Medeiros, A. A. 2006. Process mining with the heuristics miner-algorithm. Technische Universiteit Eindhoven, Tech. Rep. WP, 166.

[16] Witten, I.H., Eibe, F., Hall, M.A. 2001. Data Mining, Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufman Publishers.

ARTÍCULO 5

Referencia:

- Bogarín, A., Romero, C., & Cerezo, R. (2015). Discovering Students' Navigation Paths in Moodle. In: *Educational Data Mining (EDM).* Madrid, Spain. 556-557.

Medidas de Calidad Científica:

- CORE B
- Educational Data Mining
- Número total de citas en GoogleShoolar: 1

Discovering students' navigation paths in Moodle

Alejandro Bogarín
Department of Computer Science
University of Cordoba, Spain
(+34) 679 30 54 86
abogarin@uco.es

Cristóbal Romero
Department of Computer Science
University of Cordoba, Spain
(+34) 653 46 28 13
cromero@uco.es

Rebeca Cerezo
Department of Psychology
University of Oviedo, Spain
(+34) 627 60 70 21
cerezorebeca@uniovi.es

ABSTRACT
In this paper, we apply clustering and process mining techniques to discover students' navigation paths or trails in Moodle. We use data from 84 undergraduate Psychology students who followed an online course. Firstly, we group students using on Moodle's usage data and the students' final grades obtained in the course. Then, we apply process mining with each cluster/group of students separately in order to obtain more specific and accurate trails than using all logs together.

Keywords
Clustering, process mining, navigation paths, trails in education.

1. INTRODUCTION
One of the current promising techniques in EDM (Educational Data Mining) is Educational Process Mining (EPM). The main goal of EPM is to extract knowledge from event logs recorded by an educational system [4]. It have been observed that students show difficulties when learn in hypermedia and Computer Based Learning Environments (CBLEs) due to these environments seems to be highly cognitive and metacognitive demanding [1]. In this sense, the models discovered by EPM could be used: to get a better understanding of the underlying educational processes, to early detect learning difficulties and generate recommendations to students, to help students with specific learning disabilities, to provide feedback to either students, teachers or researchers, to improve management of learning objects, etc. In a previous work [2], we found two problems when using EPM: 1) the model obtained could not fit well to the general students' behaviour and 2) the model obtained could be too large and complex to be useful for a student or teacher. In order to solve these problems, we proposed to use clustering to improve both the fitness and comprehensibility of the obtained models by EPM. However, in this paper we propose to use a Hypertext Probabilistic Grammar (HPG) algorithm instead of Heuristics Net [2] because it provides more informative graphs.

2. METHODOLOGY
A traditional approach would use all event log data to reveal a process model of student's behaviour . Nevertheless, in this paper, we propose an approach that uses clustering for improving EPM (see Figure 1). The proposed approach firstly applies clustering in order to group students with similar features. And then, it applies process mining for discovering more accurate models of students' navigation paths or trails. In fact, we propose to use two different grouping methods:

1) Clustering students directly by using the students' grades obtained in the final exam of the course.
2) Clustering students by using a clustering algorithm over the student's interaction with the Moodle's course.

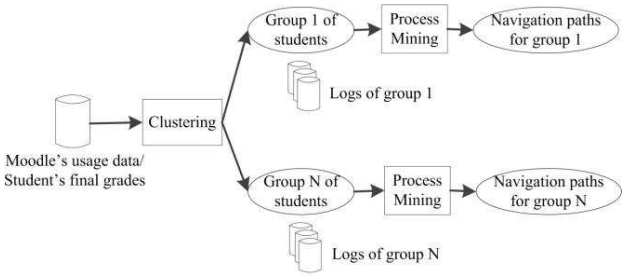


Figure 1: Proposed approach for discovering students' navigation paths.

3. DESCRIPTION OF THE DATA AND EXPERIMENTS
In this work we have used real data collected from 84 undergraduate Psychology students who followed a Moodle course. Firstly, we have divided the student's log provided by Moodle in two different ways. In a first way, we divided directly the original log file into two datasets: one that contains the 68 students who passed the course and other with the 16 students who failed. In the second way, we have used the Expectation-Maximization (EM) clustering algorithm provided by Weka [6] in order to group together students of similar behaviour when using Moodle. In this case we have obtained three clusters/datasets with the following distribution:

- **Cluster 0**: 23 students (22 pass and 1 fail).
- **Cluster 1:** 41 students (39 pass and 2 fail).
- **Cluster 2**: 20 students (13 fail and 7 pass).

After clustering, we applied EPM through HPG over the previous datasets. We have used the HPG model in order to efficiently mine trails or navigation paths [3]. HPG uses a one-to-one mapping between the sets of non-terminal and terminal symbols. Each non-terminal symbol corresponds to a link between Web pages. Moreover, there are two additional artificial states, called *S* and *F*, which represent the start and finish states of the navigation sessions respectively. The probability of a grammar string is given by the product of the probability of the productions used in its derivation. The number of times a page was requested, and the number of times it was the first and the last page (state) in a session, can easily be obtained from the collection of student navigation sessions. The number of times a sequence of two pages appears in the sessions gives the number of times the corresponding link was traversed. The aim is to identify the subset of these trails that correspond to the rules that best characterize the student's behavior when visiting the Moodle course. A trail is included only if its derivation probability is above a cut-point.

The cut-point is composed of two distinct thresholds (support and confidentiality). The support (Sup) value is for pruning out the strings whose first derivation step has low probability, corresponding to a subset of the hypertext system rarely visited. The confidence (Con) value is used to prune out strings whose derivation contains transitive productions with small probabilities. Support and confidence thresholds give the user control over the quantity and quality of the obtained trails, while α (Alp) modifies the weight of the first node in a student navigation session: when α is near 0, only those routes that start in a node which started a session are generated; when α is near 1, all weights are completely independent of the order within the session.

4. RESULTS

We have carried out several experiments with the HPG algorithm to test several configurations of number of Nodes, Links, Routes, and average route length (Avg). Results obtained when using different datasets and parameters are displayed in Table 1.

Table 1. Results with different datasets and configurations.

Dataset	Alp	Sup	Con	Nodes	Links	Routes	Avg
Fail	0,2	0,05	0,5	8	7	12	3,85
Pass	0,2	0,05	0,5	12	11	20	3,81
Cluster0	0,2	0,05	0,5	8	6	12	4,16
Cluster1	0,2	0,05	0,5	9	7	14	4,14
Cluster2	0,2	0,05	0,5	5	4	6	2,75
Fail	0,4	0,06	0,3	15	15	27	3,8
Pass	0,4	0,06	0,3	25	27	47	3,96
Cluster0	0,4	0,06	0,3	13	12	21	3,66
Cluster1	0,4	0,06	0,3	15	17	29	4,11
Cluster2	0,4	0,06	0,3	12	9	18	3,66
Fail	0,5	0,06	0,3	20	19	36	4
Pass	0,5	0,06	0,3	37	41	72	4,07
Cluster0	0,5	0,06	0,3	19	17	31	3,7
Cluster1	0,5	0,06	0,3	20	21	38	4,19
Cluster2	0,5	0,06	0,3	12	9	18	3,66

Table 1 show that the smaller and more comprehensible models were obtained using logs from students who failed (Fail dataset) and students of Cluster 2. On the other hand, the models obtained with the other datasets were much bigger and complex. We think that this may be due to:

- Both dataset Fail and Cluster 2 contain mainly information about bad students who failed the course. This type of students has a low interaction with Moodle and so, they show only some frequent navigation paths.
- Datasets Pass, Cluster0 and Cluster1 contain mainly information about good students who pass the course. This type of students has a high interaction with Moodle and so, they show more frequent navigation paths.

Finally, we show an example of obtained model when using the Cluster2 dataset. In Figure 2, each node represents a Moodle's Web page, and the directed edges (arrows) indicate how the students have moved between them. These paths can be stochastically modeled as Markov chains [5] on the graph, where the probability of moving from one node to another is determined by which Web page the student is currently visiting. Edge thickness varies according to edge weight; this allows the learning designer to quickly focus on the most important edges, ignoring those that have very low weights. In addition, line widths and numerical weights are also available.

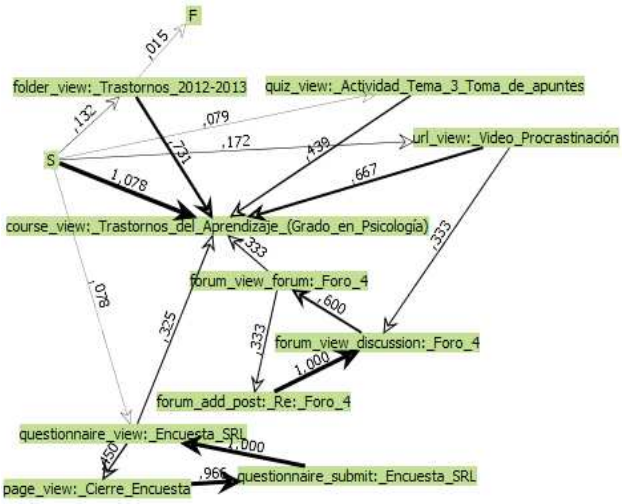


Figure 2: Navigation paths of Cluster 2 students.

Starting from Figure 2 we can see and detect what are the most frequent actions (view forum X, view questionnaire Y, view quiz Z, etc.) and in which order (navigation paths or trails) were done/followed by Cluster 2 students (normally fail students).

5. REFERENCES

[1] Azevedo, R., Behnagh, R., Duffy, M., Harley, J. M., & Trevors G. J. Metacognition and self-regulated learning in student-centered learning environments. *Theoretical foundations of student-center learning environments*. Erlbaum, Mahwah, NJ, 2nd edition, 216–260, 2012.

[2] Bogarin, A. Romero, C., Cerezo, R., Sanchez, M. Clustering for improving Educational Process Mining. *Learning Analytics and Knowledge Conference*, Indianapolis, 11-14.

[3] Borges, J., Levene, M. Data Mining of user navigation patterns. Proc. of Workshop Web Usage Analysis and User Profiling. San Diego, 2000. pp. 31-36.

[4] Pechenizkiy, M., Trcka, N., Vasilyeva, E., van der Aalst, W.M., & De Bra, P. 2009. Process Mining Online Assessment. Data. *Educational Dada Mining Conference*, Cordoba, Spain, 279-288.

[5] Kemeny, J.G., Snell, J.L. Finite Markov chains. Princeton: Van Nostrand. 1960

[6] Witten, I.H., Eibe, F., Hall, M.A. Data Mining, Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufman Publishers, 2001.

ARTÍCULO 6

Referencia:

- Bogarín, A., Romero, C. y Cerezo, Rebeca. (2015). Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle. In: *EDMETIC*, *5*(1), 73-92.

Medidas de Calidad Científica:

- Revista Nacional
- Educación Mediática y TIC

edmetic
Revista de Educación Mediática y TIC

Aplicando minería de datos para descubrir rutas de aprendizaje frecuentes en Moodle

Applying data mining to discover common learning routes in Moodle

Fecha de recepción: 10/12/2014
Fecha de revisión: 21/05/2015
Fecha de aceptación: 04/10/2015

APLICANDO MINERÍA DE DATOS PARA DESCUBRIR RUTAS DE APRENDIZAJE FRECUENTES EN MOODLE

APPYING DATA MINING TO DISCOVER COMMON LEARNING ROUTES IN MOODLE

Alejandro Bogarín Vega[1], Cristóbal Romero Morales[2] & Rebeca Cerezo Menéndez[3]

Resumen:

En este artículo, aplicamos técnicas de minería de datos para descubrir rutas de aprendizaje frecuentes. Hemos utilizado datos de 84 estudiantes universitarios, seguidos en un curso online usando Moodle 2.0. Proponemos agrupar a los estudiantes, en primer lugar, a partir de los datos de una síntesis de uso de Moodle y/o las calificaciones finales de los alumnos en un curso. Luego, usamos los datos de los logs de Moodle sobre cada cluster/grupo de estudiantes separadamente con el fin de poder obtener más específicos y precisos modelos de procesos del comportamiento de los estudiantes.

Palabras claves:

Base de datos; aprendizaje; estudiante; red de información.

Abstract:

In this paper, we apply techniques data mining to discover common learning routes. We have used data from 84 undergraduate college students who followed an online course using Moodle 2.0. We propose to group students firstly starting from data about Moodle's usage summary and/or the students' final marks in the course. Then, we use data from Moodle's logs about each cluster/group of students separately in order to be able to obtain more specific

[1] Universidad de Córdoba. abogarin@uco.es
[2] Universidad de Córdoba. cromero@uco.es
[3] Universidad de Oviedo. cerezorebeca@gmail.com

and accurate process models of students' behaviour.

Keywords:

Database; learning; student; information network.

1. Introducción

Desde la aparición de las plataformas e-learning (Moodle, WebCT, Claroline, etc.) y el modo de aprendizaje virtual que ello conlleva, las técnicas de minería de datos están siendo bastante utilizadas en la educación. Los sistemas de información almacenan todas las actividades en ficheros o bases de datos que, procesados correctamente, pueden ofrecer información muy relevante para el profesor. Por ejemplo, un profesor puede saber el comportamiento que tienen los estudiantes en la plataforma y descubrir el proceso de aprendizaje que llevan a cabo. Con esto, un profesor podrá adaptar sus cursos al modo en que trabajan sus alumnos y tomar medidas ante los problemas que se puedan detectar. Es decir, esta información útil que recopilan los sistemas información educativos puede utilizarse para tomar decisiones y responder a preguntas, buscando la mejora de la calidad y la rentabilidad del sistema educativo.

El nuevo conocimiento descubierto por las técnicas de minería de datos sobre sistemas de información e-learning es una de las áreas que aborda Educational Data Mining (EDM) (Romero, Ventura y García, 2008). Este nuevo conocimiento, puede ser útil tanto para los profesores como para los estudiantes. A los estudiantes se les puede recomendar actividades y recursos que favorezcan su aprendizaje, y los profesores, pueden obtener una retroalimentación objetiva para su enseñanza. Los profesores pueden evaluar la estructura del curso y su eficacia en el proceso de aprendizaje, y también, clasificar a los alumnos en grupos en función de sus necesidades de orientación y seguimiento.

Process Mining (PM) (Trcka y Pechenizkiy, 2009) es una técnica para hacer minería de datos sobre las aplicaciones que generan registro de eventos para identificar posibles procesos en una variedad de dominios de aplicación. La aplicación de las actividades de la minería de procesos debe tener como resultado modelos de flujos de procesos de negocio y de información de su empleo histórico (caminos más frecuentes, actividades menos realizadas, etc.).

Herramientas de PM como ProM (VAN DER AALST, 2011) brindan análisis y descubrimiento de flujos de procesos a partir de los registros de eventos generados por muchas aplicaciones.

Este paper está organizado de la siguiente forma: El siguiente capítulo muestra la metodología utilizada, a continuación se describen los datos usados. En la sección 4 se describe los experimentos realizados y, finalmente, se muestran las conclusiones y futuras mejoras.

2. Metodología

Proponemos una metodología que utiliza clustering para agrupar a los alumnos por tipos y así poder mejorar los modelos extraídos con minería de procesos.

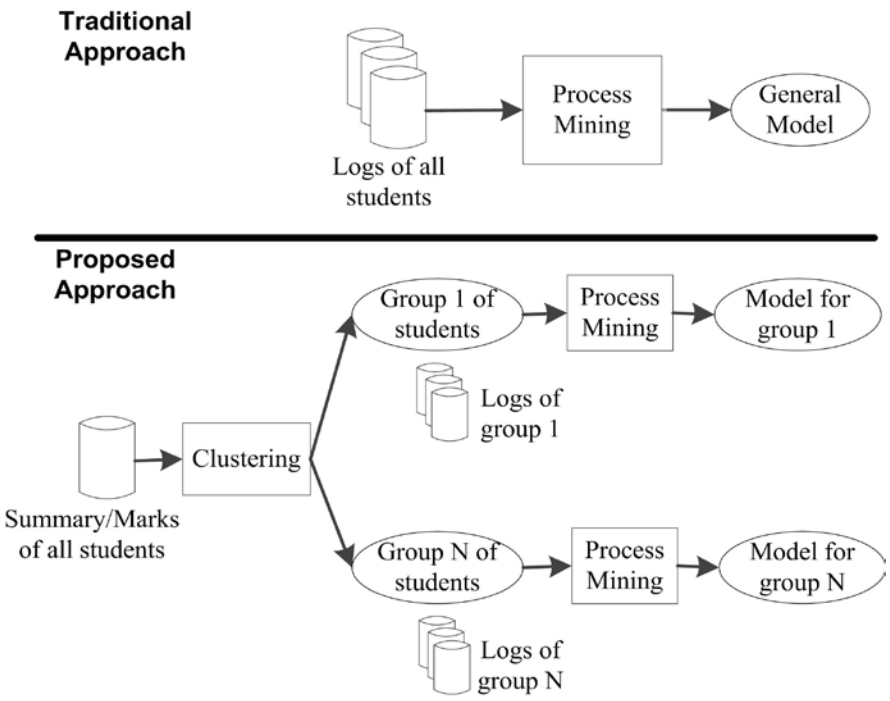


Figura 1: Investigación Tradicional VS Investigación Propuesta.

Fuente: Elaboración propia.

Las técnicas tradicionales de investigación en EDM y PM utilizan todos los datos de los registros de eventos para descubrir un modelo general de proceso del comportamiento de todos los estudiantes.

En cambio, nuestra propuesta aplica agrupamiento previo precisamente para obtener grupos de estudiantes con similares características.

Posteriormente, se aplica minería de procesos para descubrir modelos específicos de los comportamientos de los estudiantes. Se proponen, como se muestra en la figura 1, dos tipos diferentes de clustering/agrupamiento:

- Manual: Se agrupan a los estudiantes directamente usando la nota final obtenida en el curso
- Automática: Se agrupan a los estudiantes aplicando clustering sobre la información de interacción que éstos realizan al ejecutar el curso en la plataforma Moodle

En el agrupamiento Manual nos encontramos con dos tipos de alumnos:

1. Alumnos cuya nota final es menor a 5 (alumnos suspensos)
2. Alumnos cuya nota final es mayor o igual a 5 (alumnos aprobados)

Para la agrupación Automática, las variables utilizadas y su descripción para realizar el clustering provienen de la interacción que los estudiantes realizan en Moodle, y son las que se pueden ver en la tabla 1.

Estas variables tienen un valor determinado para cada uno de los alumnos que se estudian en este trabajo. Según los valores de interacción que presenten los estudiantes, se les asociará con uno de los tres clusters de nuestro estudio.

3. Descripción de los datos usados

Los datos utilizados en este estudio fueron obtenidos de un curso de Moodle 2.0 utilizado por 84 estudiantes universitarios del grado de Psicología de la Universidad de Oviedo. El estudio se realizó durante el curso académico 2012-2013. La investigación se realiza sobre una asignatura de carácter obligatorio de tercero de carrera.

El profesor pidió a los estudiantes que participasen en un programa de

e-Learning denominado "aprendiendo a aprender", relacionado con la temática de la asignatura y que se completaba en horario fuera de clase. El programa se compone de 11 unidades diferentes que se mandaban a los estudiantes semanalmente, pero cada uno de ellos podía trabajar en la unidad durante un periodo de 15 días.

Cada unidad se compone de tres tipos de contenidos:

- Nivel de conocimiento declarativo: contenidos teóricos, de información y de cómo poner la estrategia o estrategias semanal de "aprender a aprender" en práctica
- Nivel de conocimiento procedimental: tareas prácticas donde los estudiantes tienen que poner en práctica su conocimiento declarativo
- Nivel de conocimiento condicional: foros de discusión donde los estudiantes tienen que tratar temas de cómo tendrían o podrían usar la estrategia o estrategias de la semana en diferentes contextos

Los estudiantes consiguen un punto extra en su calificación final de la asignatura si completan al menos el 80 % de las tareas.

Las tareas obligatorias de cada unidad eran: realizar la tarea práctica y publicar, al menos, un comentario en cada foro.

Las tareas sugeridas de cada unidad eran: comprender los contenidos teóricos y ponerlos en práctica en la tarea y compartir su experiencia sobre el tema de la semana en el foro.

Se utilizan varias fuentes de información diferentes en las que se basan los datos obtenidos del trabajo realizado por los estudiantes en todo el programa.

Por un lado, se muestra en la tabla 1 las variables que se tienen en cuenta, que determina la interacción que tiene cada estudiante en la plataforma Moodle. Estas variables se calculan a partir del registro de Moodle y diferentes tablas de bases de datos.

Tabla 1: Variables que muestran la interacción de los estudiantes en Moodle.

Fuente: Elaboración propia.

Nombre	Descripción	Método de Extracción
Tiempo de Teoría	Tiempo total empleado en componentes teóricos de los contenidos	La suma de los periodos entre resource view y la próxima acción diferente
Tiempo de Tareas	Tiempo total empleado en tareas de enseñanza	La suma de los periodos entre quiz view/quiz attempt/quiz continue attempt/quiz close attempt y la próxima acción diferente
Tiempo de Foros	Tiempo total empleado en la revisión de foros	La suma de los periodos entre forum view y la próxima acción diferente
Días de Teoría	Cuantos días, en un periodo de 15 días, esperan para comprobar el contenido al menos una vez (en días)	Fecha de resource view desde que el contenido está disponible
Días de Tareas	Cuantos días, en un periodo de 15 días, esperan para comprobar la tarea al menos una vez (en días)	Fecha de task view desde que la tarea está disponible
Días de "entrega"	Cuantos días, en un periodo de 15 días, tardan en completarlas (en días)	Fecha de quiz close attempt desde que la tarea está disponible
Palabras en los Foros	Número de palabras publicadas en foros	Extraer el número de palabras de todo lo publicado de forum add discussion OR forum add replay
Frases en los Foros	Número de frases publicadas en foros	Etraer el número de frases de todo lo publicado de forum add discussion OR forum add replay

Estos datos obtenidos de Moodle y las diferentes tablas de bases de datos son procesados y se convierten en un fichero .ARFF al que se le aplicará posteriormente agrupamiento manual o un algoritmo de clustering proporcionado por el software de DM WEKA (Witten y Frank, 2005).

Por otro lado, se ha usado también el fichero de registro proporcionado

por Moodle con los campos que se muestran en la tabla 2.

Tabla 2: Variables del registro de eventos (LOG) de Moodle.

Fuente: Elaboración propia.

Atributos	Descripción
Curso	El nombre del curso
Dirección IP	La IP del dispositivo usado para acceder
Tiempo	La fecha de acceso
Nombre Completo	El nombre del estudiante
Acción	La acción que realiza el estudiante
Información	Más información sobre la acción

De este fichero nos quedamos con cuatro variables, ya que, no utilizamos la variable Curso (todos los registros tienen el mismo valor) y Dirección IP (para nuestro propósito es una información irrelevante). También hemos sustituido el nombre de los estudiantes por Ids (Identificadores) para mantener su privacidad, y hemos filtramos las acciones de nuestro fichero log.

Además, de 39 posibles acciones que almacena Moodle solo hemos usado las 20 acciones que son relevantes para el rendimiento de los estudiantes durante el curso: assignment upload, assignment view, course view, folder view, forum add discusión, forum add post, forum update post, forum view discusión, forum view forum, page view, questionnaire submit, questionnaire view, quiz attempt, quiz close attempt, quiz continue attempt, quiz review, quiz view, quiz view summary, resource view y url view.

Se considera que las acciones como ver todos los usuarios, ver todas las etiquetas, ver todas las carpetas, etc., no tienen relevancia en la calificación final.

El filtrado que se realiza tiene bastante sentido ya que, el fichero original pasa de tener 41532 registros a 40466, es decir, se eliminan muy pocos registros, lo que indica que estas acciones no resultaban significativas en el rendimiento final de los estudiantes.

Es importante comentar que el campo información contiene

información adicional sobre las acciones que se realizan en la plataforma Moodle. Por ejemplo, una determina acción como quiz view tiene asociado 25 campos con informaciones diferentes.

En total hay 332 eventos (acciones más el campo información) que pueden realizar los estudiantes y los que se consideran a la hora de realizar la experimentación y extraer los resultados.

Finalmente, transformamos los ficheros obtenidos a formato MXML (Minimal XML) usando ProMimport framework para que pueda ser interpretado por ProM (Van Der Aalst, 2011), obteniendo seis conjuntos de datos sobre los que realizamos experimentación:

- Todos los estudiantes (84 estudiantes)
- Estudiantes que aprueban (68 estudiantes)
- Estudiantes que suspenden (16 estudiantes)
- Estudiantes que pertenecen al cluster 0 (22 aprueban y 1 suspenden)
- Estudiantes que pertenecen al cluster 1 (39 aprueban y 2 suspenden)
- Estudiantes que pertenecen al cluster 2 (13 aprueban y 7 suspenden)

4. Resultados de Experimentación

Se han realizado varios experimentos para probar nuestra propuesta. En el primero, se utilizaron todos los datos del registro de los 84 estudiantes. En el segundo, se dividió el archivo de registro original en dos conjuntos de datos: una que contiene 68 estudiantes que aprobaron y otro con 16 estudiantes que suspendieron. En el último experimento, se ha utilizado el algoritmo de clustering proporcionada por Weka (Witten y Frank, 2005) Esperanza-Maximización (EM) para agrupar alumnos de similares características utilizando las variables que aparecen en la tabla 1. Se utilizó este algoritmo por ser un algoritmo de clustering bien conocido y además, no requiere que el usuario especifique el número de grupos. En nuestro caso, se obtuvieron tres grupos con la siguiente distribución de los alumnos:

- Cluster 0: 23 estudiantes (22 aprueban y 1 suspenden)
- Cluster 1: 41 estudiantes (39 aprueban y 2 suspenden)
- Cluster 2: 20 estudiantes (13 suspenden y 7 aprueban)

Hemos utilizado la herramienta de código abierto ProM (Van Der Aalst, 2011), que es un software específico para temas relacionados con la minería de procesos y hemos aplicado el algoritmo Heuristics Miner que está basado en la frecuencia de patrones, debido a que concentra su comportamiento principal en el registro de eventos.

Asimismo, el Heuristic Miner es una red heurística dibujada como un grafo cíclico dirigido, el cual muestra, en nuestro caso, el comportamiento más frecuente de los estudiantes en cada conjunto de datos utilizados.

Se usa los parámetros por defecto del algoritmo Heuristic Miner de ProM (Van Der Aalst, 2011) y como medida de calidad, el Ajuste o Fitness.

El Ajuste indica la diferencia entre el comportamiento realmente observado en el registro y el comportamiento descrito por el modelo de proceso. Una secuencia de actividades que pertenecen a un mismo caso se llama traza. Las trazas del registro pueden estar asociadas con rutas de ejecución especificadas por el modelo de proceso. Si el modelo tiene un valor de Ajuste bajo, indica que el modelo de minería de procesos no analiza correctamente la mayoría de las trazas de registro. Esto puede ser debido a la presencia de ruido, resultado de actividades que no se tienen en cuenta y conexiones que faltan.

Tabla 3: Resultados del valor de Ajuste de los diferentes modelos.

Fuente: Elaboración propia.

Conjunto de Datos	Ajuste
Todos los Estudiantes	0.8333
Aprobados	0.9117
Suspensos	0.9375
Estudiantes Cluster 0	0.9130
Estudiantes Cluster 1	0.9024

Estudiantes Cluster 2	0.9000

Se puede ver en la tabla 3, que el valor más bajo de la medida de Ajuste se obtuvo cuando se utilizó todos los datos de los estudiantes conjuntamente, en los que 70 de los 84 estudiantes encajan con el modelo obtenido, es decir, el 83,33 % de todos los estudiantes. Por otro lado, todos los otros modelos (obtenido usando clustering tanto de forma manual como automática) obtienen un valor de Ajuste superior al 90 % en todos los casos. El mayor valor de Ajuste, se obtuvo cuando se usó los datos de los estudiantes que suspendían, donde 15 de los 16 estudiantes encajan en el modelo obtenido, es decir, el 93,75 % de los estudiantes que suspenden. Por lo tanto, en este caso se puede ver que estos modelos específicos obtenidos usando clustering manual y automático representan/encajan mejor que el modelo general obtenido de todos los estudiantes.

En la tabla 4, se muestra información sobre el nivel de complejidad o tamaño de cada una de los modelos obtenidos.

Se han usado dos medidas típicas de la teoría de grafos (el número total de nodos y el número total de enlaces) con el fin de ver el nivel de complejidad de los modelos obtenidos.

Tabla 4: Complejidad/Tamaño de los modelos obtenidos.

Fuente: Elaboración propia.

Conjunto de Datos	**N.Nodos**	**N.Enlaces**
Todos los Estudiantes	32	70
Estudiantes aprobados	113	244
Estudiantes Suspensos	12	24
Estudiantes Cluster 0	61	121
Estudiantes Cluster 1	59	110
Estudiantes Cluster 2	38	84

Se puede ver en la tabla 4, que el modelo más pequeño y por tanto más fácilmente comprensible se obtuvo con los estudiantes suspensos, seguido

por todos los estudiantes y, los estudiantes del cluster 2. Por otro lado, los otros tres modelos son mucho mayores y complejos. Se cree que las razones podrían ser:

- En el conjunto de datos de todos los estudiantes, los estudiantes muestran diferentes comportamientos y sólo tienen algunas acciones en común porque hay mezclados diferentes tipos de estudiantes (aprobados y suspensos).
- En el conjunto de datos de los estudiantes que suspenden y del cluster 2, los estudiantes muestran sólo algunos patrones de comportamiento común porque este tipo de estudiantes participa/interactúa poco con la plataforma Moodle.
- En el conjunto de datos de los estudiantes que aprueban, cluster 0 y cluster 1, los estudiantes muestran muchos mas patrones de comportamiento comunes porque este tipo de estudiantes son usuarios más activos de Moodle.

Finalmente, se muestran los modelos que obtienen el mejor y peor ajuste. En el primer ejemplo se muestra la red heurística obtenida cuando se usan todos los alumnos y en el segundo la de los estudiantes que suspenden.

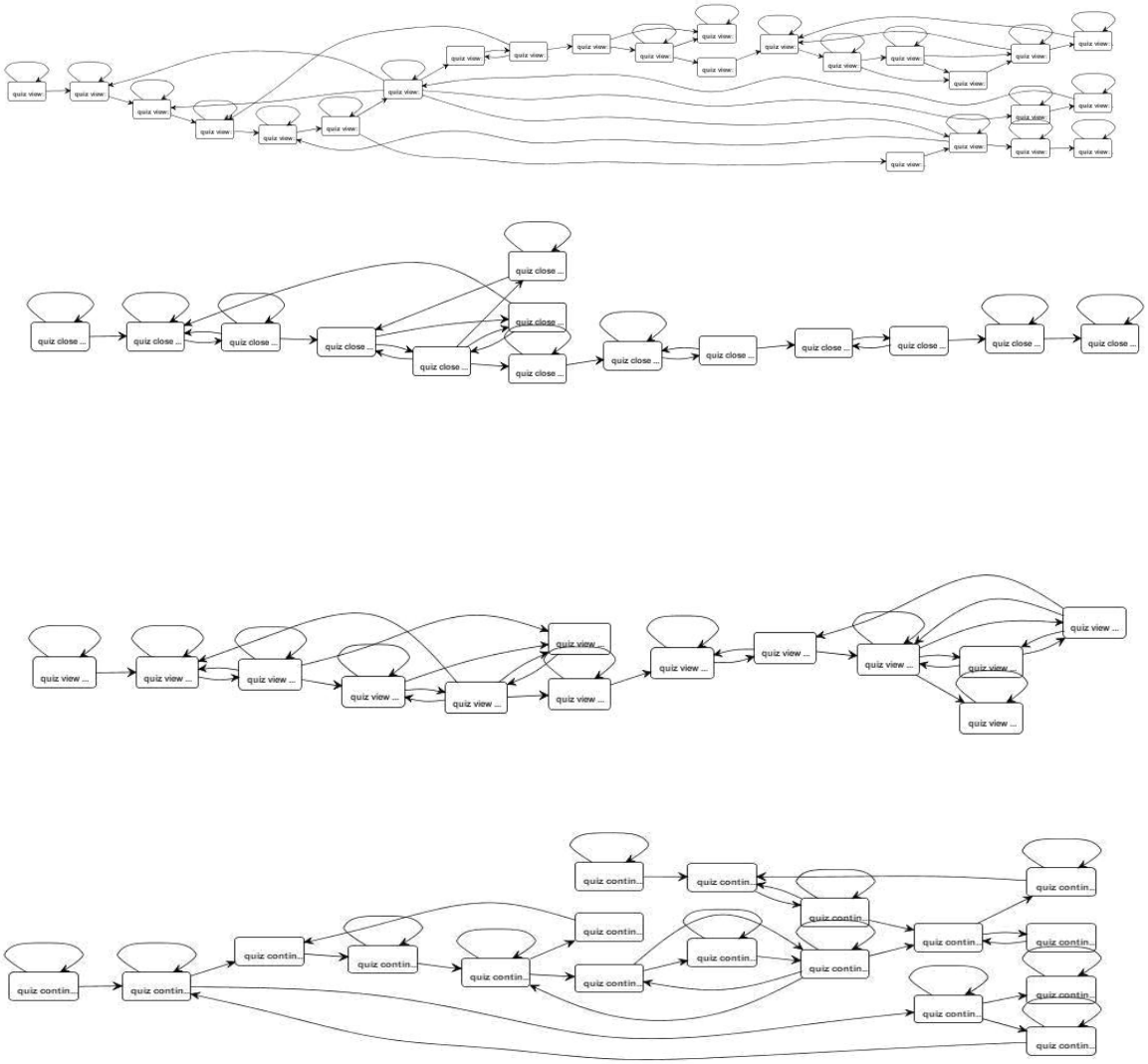
En nuestras redes heurísticas las cajas representan los eventos realizados por lo estudiantes cuando interactúan con la plataforma Moodle y los arcos/enlaces representan las relaciones/dependencias entre los eventos.

En la figura 2 se pueden ver dos subredes que siguen la mayoría de los estudiantes del curso en estudio. La mayor subred consta de algunos eventos de ver en el foro de todos los foros que más se han visto en el curso. Y la menor subred contiene algunos eventos de ver exámenes de los exámenes más vistos en el curso.

Figura 2: Red Heurística de todos los estudiantes.
Fuente: Elaboración propia.

A continuación, en la figura 3 se muestra la red heurística obtenida cuando se usan los alumnos que aprueban en el curso.

Se puede ver que estos alumnos tienen un mayor número de subredes asociadas debido a que la interacción con la plataforma es mayor y por tanto hay una mayor diversidad en cuanto a eventos comunes entre estos alumnos.



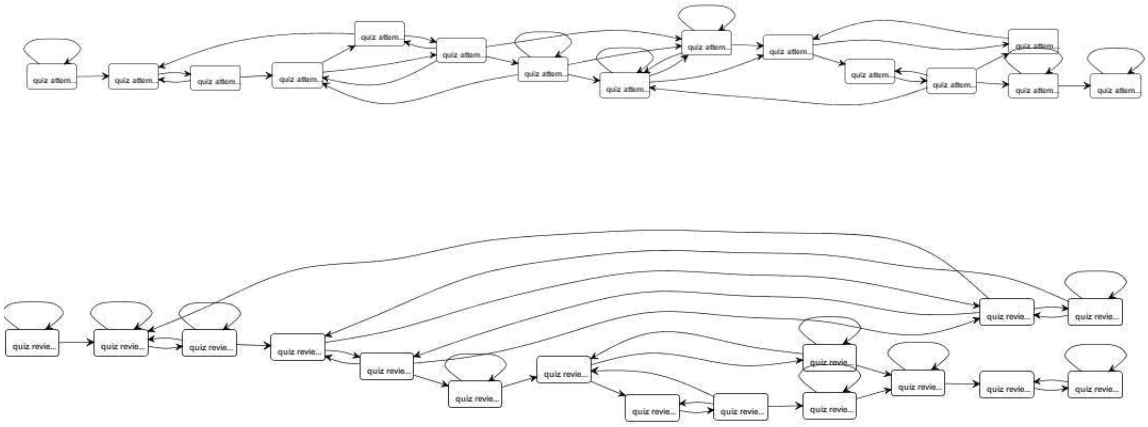


Figura 3: Red Heurística de los estudiantes que aprueban.

Fuente: Elaboración propia.

Por otro lado, la figura 4 muestra dos subredes que siguen la mayoría de los estudiantes que suspenden en el curso.

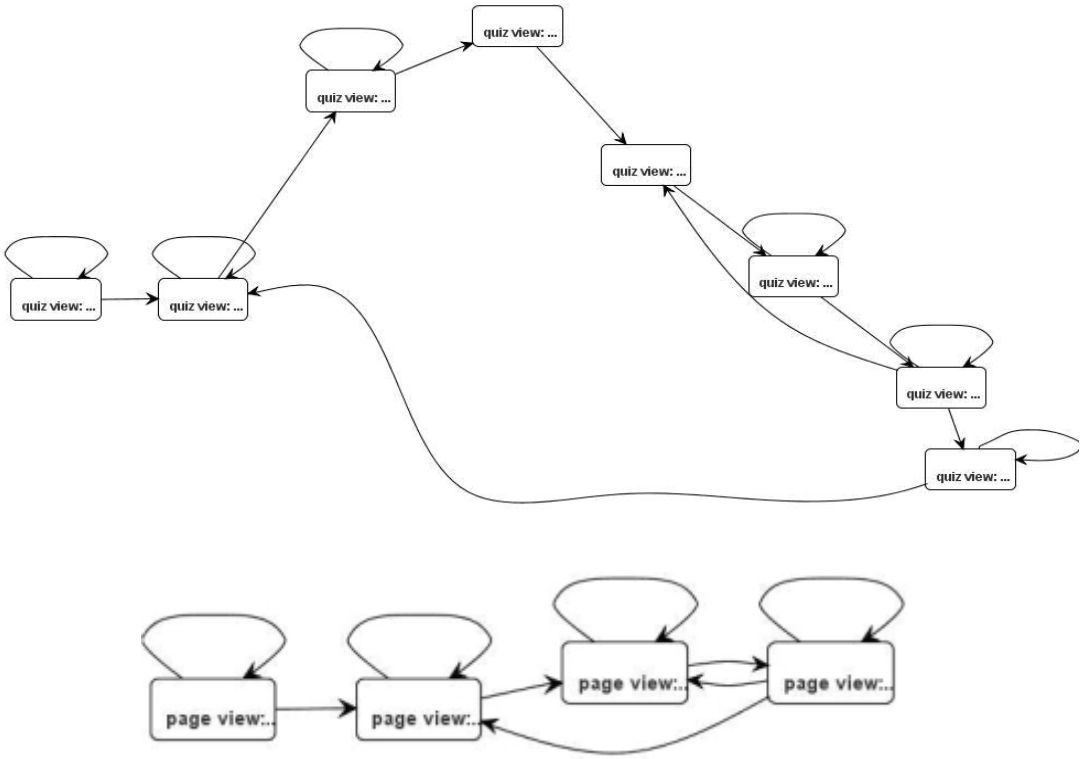


Figura 4: Red Heurística de los estudiantes que suspenden.

Fuente: Elaboración propia.

La red superior consta de algunas acciones de páginas vistas de las

páginas más visitadas. Esas páginas contienen información general sobre el curso.

La subred más pequeña contiene algunas acciones de ver exámenes de los exámenes más vistos.

En este caso/conjunto de datos, el numero de exámenes es mucho menor que en el caso de los estudiantes que aprueban y no en el mismo orden de visita.

Desde un punto de vista educativo y práctico (se podría usar esta información para proporcionar retroalimentación a los profesores sobre el aprendizaje del estudiante), podría fácilmente ser usado para señalar nuevos estudiantes con riesgo de suspender en el curso. Por ejemplo, los profesores sólo tienen que comprobar si los nuevos estudiantes siguen las mismas rutas específicas/patrones de comportamiento que muestra la red heurística de los estudiantes que suspenden. Es decir, si visitan las mismas páginas, ven los mismos exámenes y, en el mismo orden que los estudiantes que suspendieron anteriormente.

5. Conclusiones y Futuras Mejoras

En este trabajo se propone el uso de agrupamiento o clustering para mejorar la minería de procesos educativa y, al mismo tiempo, optimizar tanto el rendimiento/ajuste y comprensibilidad/tamaño del modelo obtenido. La comprensibilidad del modelo obtenido es un objetivo básico en la educación, debido a la transferencia de conocimientos básicos que ello conlleva.

Realizar gráficos, modelos o una representación visual más accesible o al menos, accesible, para los profesores y estudiantes, hacen que estos resultados sean muy útiles para el seguimiento del proceso de aprendizaje y para proporcionar una retroalimentación, siendo uno de nuestros futuros retos realizarlo en tiempo real. Además, Moodle no proporciona herramientas de visualización específicas de los datos usados por los estudiantes que permitan a los diferentes agentes del proceso de aprendizaje entender estas grandes

cantidades de datos "en bruto" y, tomen consciencia de lo que esta pasando en una educación a distancia, además de ampliar el uso de los resultados de Entornos de Aprendizaje Hipermedia Adaptativos en los que es muy útil motivar a los estudiantes o recomendarles rutas de aprendizaje, con el fin de mejorar la experiencia de aprendizaje de una manera más estratégica.

En el futuro, queremos hacer más experimentos para poner a prueba nuestra propuesta con otros tipos de cursos pertenecientes a diferentes áreas de conocimiento. También queremos explorar otras maneras de agrupar estudiantes antes de la minería de procesos. Asimismo, se propone realizar pruebas de selección dentro del conjunto de datos, sólo los eventos que tienen un determinado umbral de frecuencia que resulte más óptimo en el modelo de proceso extraído.

Referencias bibliográficas

AZEVEDO, R., BEHNAGH, R., DUFFY, M., HARLEY, J., y TREVORS, G. (2012). Metacognition and self-regulated learning in student-centered leaning environments. *Theoretical foundations of student-centered learning environments*, 171-197.

KLÖSGEN, W., y ZYTKOW, J. M. (2002). Handbook of data mining and knowledge discovery. Oxford: University Press, Inc.

MULDNER, K., BURLESON, W., VAN DE SANDE, B., y VANLEHN, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User Modeling and User-Adapted Interaction, 21*(1-2), 99-135.

PECHENIZKIY, M., TRCKA, N., VASILYEVA, E., VAN DER AALST, W., y DE BRA, P. (2009). *Process Mining Online Assessment Data*. International Working Group on Educational Data Mining.

PEDRAZA-PEREZ, R., ROMERO, C., &y VENTURA, S. (2011). *A Java desktop tool for mining Moodle data*. En Proceedings of 4th International Conference on Educational Data Mining (pp. 319-320).

PERERA, D., KAY, J., KOPRINSKA, I., YACEF, K., y ZAÏANE, O. R. (2009). *Clustering and sequential pattern mining of online collaborative learning data*. Knowledge and Data Engineering, IEEE Transactions on, 21(6), 759-772.

RABBANY, R., TAKAFFOLI, M., y ZAÏANE, O. R. (2011). *Analyzing participation of students in online courses using social network analysis techniques*. En Proceedings of educational data mining.

ROMERO, C., y VENTURA, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews. *IEEE Transactions on, 40*(6), 601-618.

ROMERO, C., VENTURA, S., y GARCÍA, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

ROMERO, C., VENTURA, S., ZAFRA, A., y BRA, P. D. (2009). Applying Web usage mining for personalizing hyperlinks in Web-based adaptive educational systems. *Computers & Education, 53*(3), 828-840.

SIEMENS, G., & D BAKER, R. S. (2012, April). *Learning analytics and educational data mining: towards communication and collaboration*. En Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254). ACM.

SOUTHAVILAY, V., YACEF, K., y CALVO, R. A. (2010, June). *Process Mining to Support Students' Collaborative Writing*. En EDM (pp. 257-266).

TRCKA, N., & PECHENIZKIY, M. (2009, November). *From local patterns to global models: Towards domain driven educational process mining*. En Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on (pp. 1114-1119). IEEE.

VAN DER AALST, W. M. (2011). Discovery, Conformance and Enhancement of Business Processes. Springer, Heidelberg.

WITTEN, I. H., y FRANK, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Cómo citar este artículo

Bogarín Vega, A., Romero Morales, C. y Cerezo Menéndez, Rebeca (2016). Aplicando minería de datos para descubrir rutas de aprendizaje frecuente en Moodle. *EDMETIC, Revista de Educación Mediática y TIC, 5*(1), 73-92.