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.

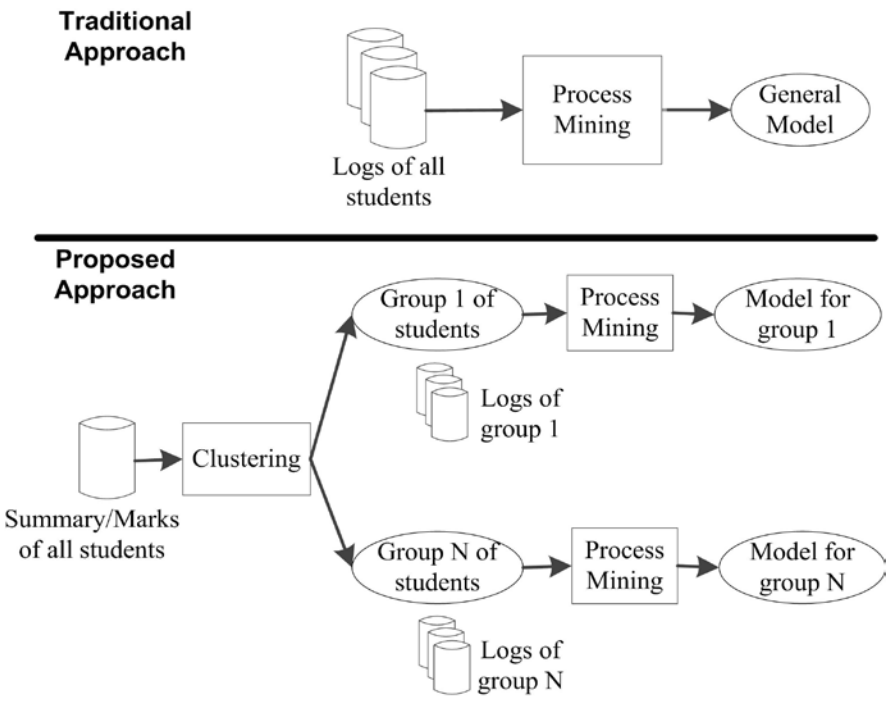


Figura 1: Investigación Tradicional VS Investigación Propuesta.

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.

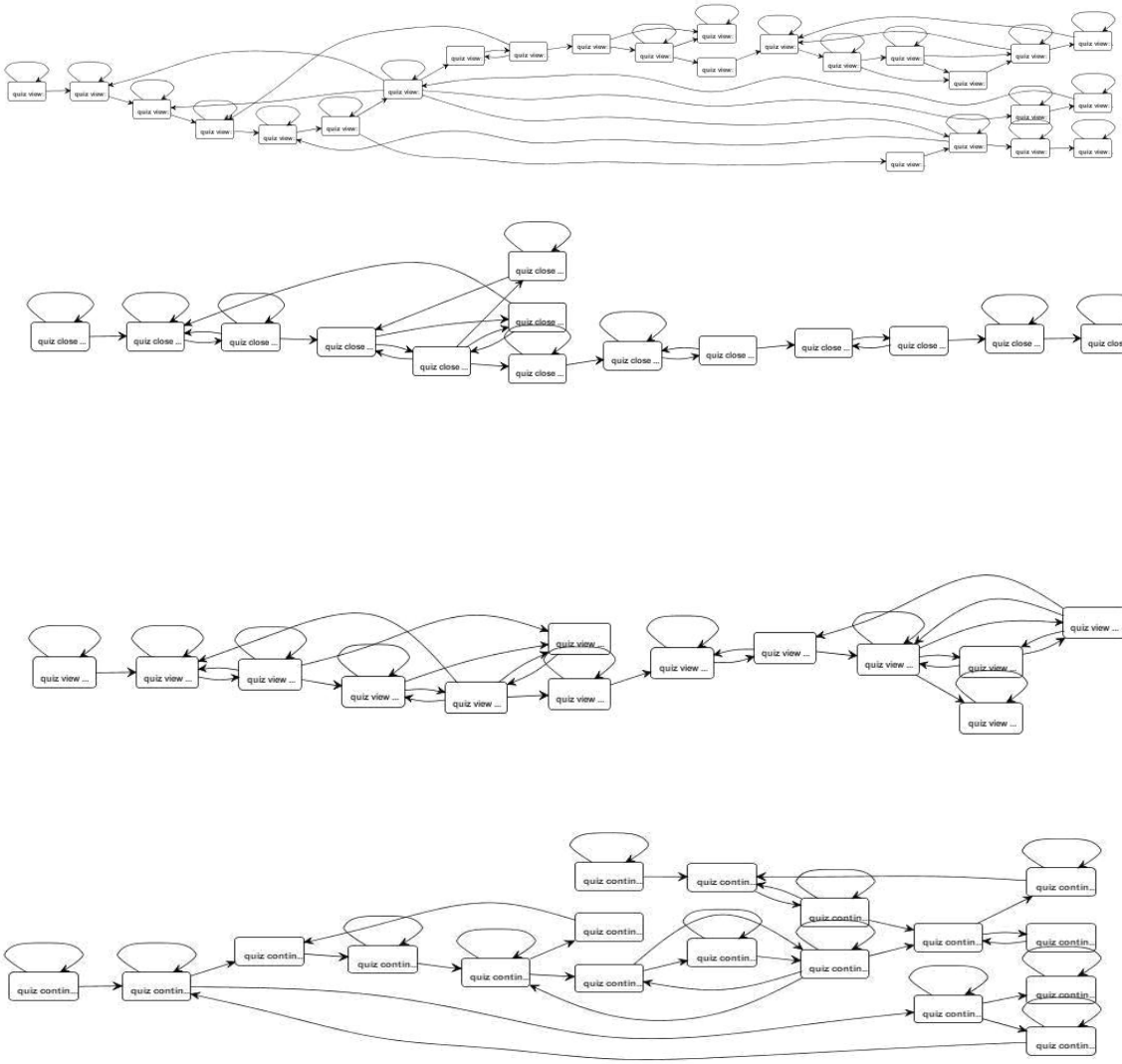
En nuestras redes heurísticas las cajas representan los eventos realizados por lo estudiantes cuando interactúan con la plataforma Moodle y los arcos/enlaces representan las relaciones/dependencias entre los eventos.

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.



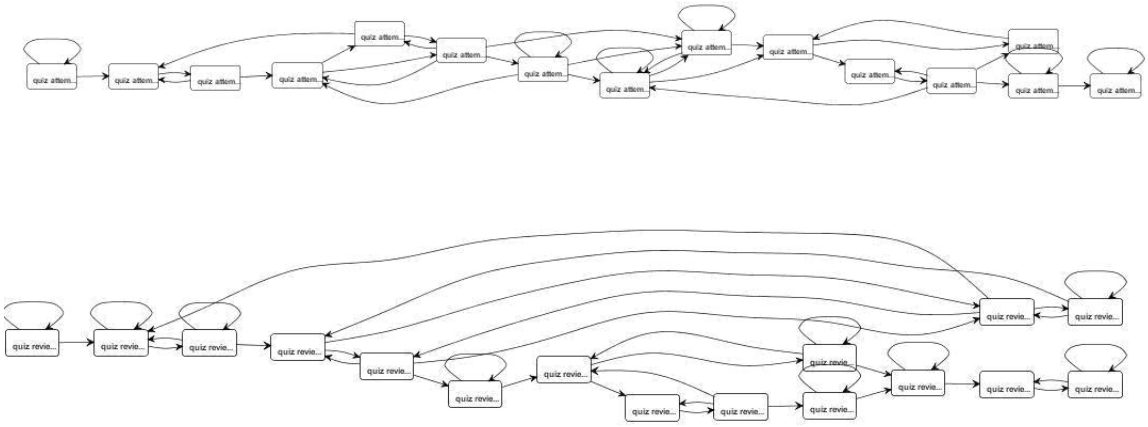


Figura 3: Red Heurística de los estudiantes que aprueban.

Fuente: Elaboración propia.

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

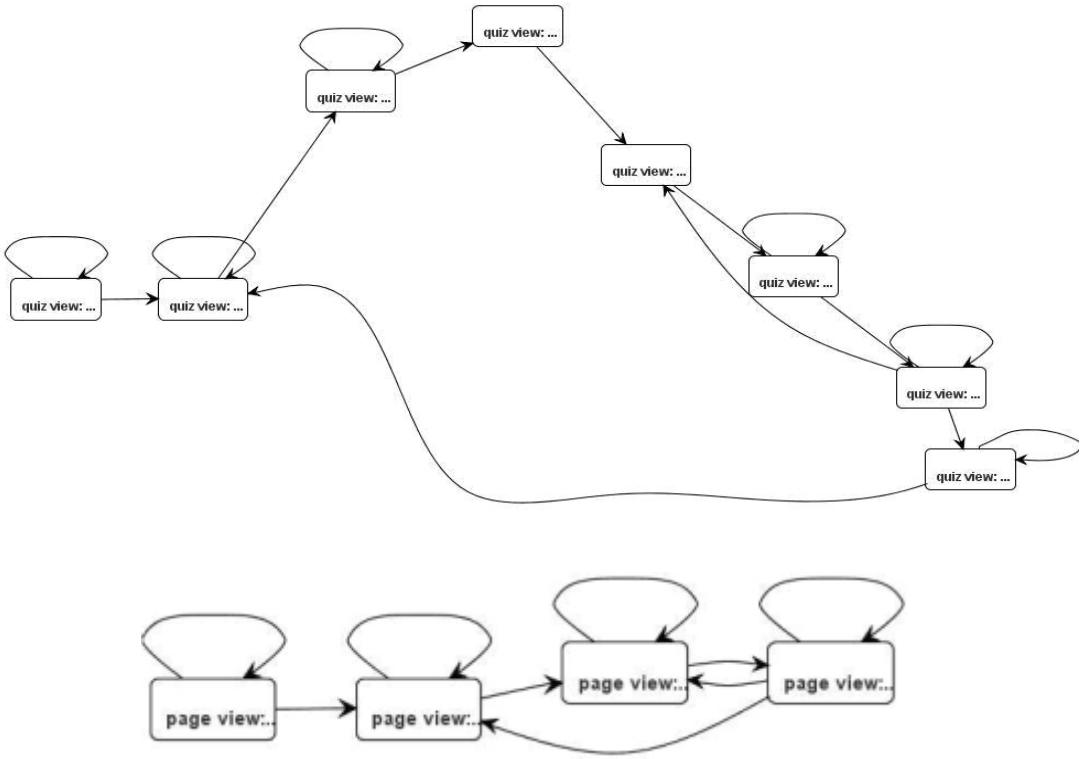


Figura 4: Red Heurística de los estudiantes que suspenden.

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.