PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

EXTRACCIÓN DE PATRONES SEMÁNTICAMENTE DISTINTOS A PARTIR DE LOS DATOS ALMACENADOS EN LA PLATAFORMA PAIDEIA

Tesis para optar el grado de Magistra en Informática con Mención en Ciencias de la Computación que presenta

NATALÍ FLORES LAFOSSE

Dirigido por

DR. HUGO ALATRISTA SALAS

Jurados

DR. HÉCTOR ANDRÉS MELGAR SASIETA

DR. HUGO ALATRISTA SALAS

MG. CLAUDIA MARÍA DEL PILAR ZAPATA DEL RÍO

San Miguel, 2016

Resumen

En la actualidad el uso de plataformas LMS (Learning Management System) se ha convertido en una necesidad en las instituciones de educación superior. Una de las plataformas más populares es Moodle, la cual se enfoca en el uso de módulos para distribuir el contenido educativo. Sin embargo, los docentes que utilizan la plataforma no suelen recibir una retroalimentación sobre el comportamiento de sus alumnos en sus cursos. Existen muchos métodos para conseguir dicha retroalimentación, encuestas o entrevistas, sin embargo el uso de los logs del sistema presenta la ventaja de almacenar información verídica del comportamiento de los usuarios.

La presente tesis busca utilizar algoritmos de Minería de Datos para extraer patrones de comportamiento semánticamente distintos de los usuarios de la plataforma, a fin de brindar retroalimentación tanto a los administradores de la plataforma como a los docentes. Se buscan patrones semánticamente distintos para así hacer un análisis con diferentes acercamientos a la misma búsqueda de información. Para ello se hace uso de la metodología *Descubrimiento de Conocimiento a partir de bases de Datos* (KDD por sus siglas en inglés), la cual establece una serie de pasos a seguir.

Aplicando dicha metodología, en principio, se realizó una selección de los datos a utilizar. A esta selección, luego, se le aplica un pre-procesamiento antes de utilizarla como entrada de los algoritmos de Minería de Datos, usando la librería SPMF y la aplicación Weka según sea el caso. Se usaron distintos algoritmos tanto para clusterizar datos, descubrir *itemsets* frecuentes y reglas de asociación y obtener patrones secuenciales.

Los resultados de clusterización resultaron en tres grupos, caracterizados por las acciones que realizan. Las reglas de asociación e *itemsets* frecuentes mostraron un comportamiento regular de los usuarios, quienes principalmente ingresan para “ver” tanto “cursos” como “recursos”. Una conclusión similar se deriva la los patrones secuenciales, los cuales repiten la acción de “ver” frecuentemente.

Finalmente, los resultados de reglas de asociación se visualizan en un grafo de fuerzas. Parte de los patrones secuenciales se usan para un grafo similar. Estos grafos junto a las figuras de clusterización sirven como resultados de los objetivos.

La tesis está dividida en seis capítulos. El primero es la introducción y contexto. Le sigue el capítulo de estado del arte y marco teórico. El capítulo 3 establece los objetivos. El capítulo 4 describe la experimentación y resultados. En el capítulo 5 se analizan y discuten los datos recabados de la experimentación. Finalmente, en el capítulo 6 se presentan las conclusiones, limitaciones del estudio y trabajos futuros.

Abstract

Nowadays, using an LMS (Learning Management System) has become a necessity in Higher Education. One of the most popular LMS is Moodle, which uses “modules” as a way of building a course. However, the teachers who use the platform, normally lack feedback about the behavior of students in their courses. Many ways to gets feedback exist, like surveys or interviews, however, using the system logs present the advantage of storing real information of user behavior.

This thesis uses data mining algorithms to extract semantically different patterns of behavior of the platform’s users, in order to bring feedback to teachers and the system’s administrators. The patterns must be semantically different, so the analysis may be approached in a more global way. It uses the *Knowledge Discovery in Databases* (KDD) methodology, which establishes a sequence of steps to follow.

Using that methodology, a selection of data was made. This selection was then subject of a preprocessing before using data mining algorithm, for which Weka and the SPMF library were applied. Different algorithms were used for clustering, discovering frequent itemsets and association rules and extracting frequent sequential patterns.

The clustering resulted in three groups, characterized by the actions users made. The association rules and frequent itemsets showed a regular behavior in users, who mostly enter the platform just to “view” “courses” and “resources”. A similar conclusion comes from the sequential patterns, which repeat the “view” action several times.

Finally, the association rules results are shown in a force graph. Also, a part of the sequential patterns results are used to build a similar graph. Those graphs, joined by the figures in the clustering result may be used as the feedback that this thesis was seeking.

This study is divided in six chapters. The first one consists in an introduction and context. It is followed by the State of Art and Theoretical Framework. Chapter 3 establishes the Goals of the study. Chapter 4 describes the experiment and its results. In Chapter 5, the data resulted in the experiment is analyzed and discussed. At the end, Chapter 6 presents the conclusions, limitations of the study and future works.

A Mim… A Pa…
Su ejemplo y cariño es lo que me hace continuar.

A Lance…
No te olvidaré.

Agradecimientos

Deseo expresar mi agradecimiento al Ing. Albert Díaz, la Ing. Virginia Villanueva y la Ing. Jackeline Trujillo por su apoyo en el acceso y permiso de uso de los datos de la plataforma PAIDEIA, así como su ayuda en general durante la realización del estudio.

Así mismo al Mag. (c) José Barturén y al Mag (c) Dennis Cohn por ser los compañeros de las tensiones y momentos de cansancio durante el desarrollo de esta Maestría.

Un agradecimiento muy importante al Dr. Hugo Alatrista-Salas, quien tomó un tema de investigación que no era propiamente su campo, y ayudó a generar una tesis. Mil gracias.

Por otro lado, a mis padres, hermanos y a Lance por su apoyo incondicional durante los dos años de la Maestría. Gracias por no desesperarse… tanto.

A Luv. Porque con amor, las cosas avanzan siempre mejor… y todo se hace posible.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1. Generalidades

1.1. Introducción

En la actualidad, ir a clase no tiene el mismo significado que antes. La educación, al igual que el resto de las actividades humanas, ha ido cambiando según la tecnología se va inyectando más en las vidas de las personas. Si antes se usaban cuadernos, ahora algunos utilizan tabletas y si los docentes revisaban los ejercicios resueltos a mano, ahora reciben documentos en digital (VATE, 2013). Para ello, se hace vital utilizar herramientas tecnológicas que integren las necesidades del alumno en un solo espacio.

Existen muchas de estas herramientas que apoyan el proceso educativo. Hay programas de software que buscan desarrollar capacidades específicas en los estudiantes sean para ciencias o letras, y formas de comunicación alternativas que mantienen al docente y alumno conectados fuera de aula. Sin embargo, el amplio ecosistema de herramientas necesita centralizarse en un punto. Es por ello que surgen como alternativa las plataformas virtuales educativas.

Las plataformas virtuales educativas brindan ese espacio de encuentro, sea sustitutorio o complementario del aula, incluyendo a su vez, actividades de aprendizaje, comunicación y evaluación para alumnos y docentes. En la actualidad, se pueden encontrar diversas plataformas como Blackboard[1], Canvas[2], Moodle[3] y otros, cada cuál con su propio conjunto de opciones y funcionalidades.

En la Pontificia Universidad Católica del Perú (PUCP) se utiliza Paideia, una adaptación de la plataforma de código abierto Moodle, tanto para cursos virtuales como presenciales. Al estar basado en Moodle, cuenta con múltiples módulos desarrollados tanto por los que mantienen el núcleo de la plataforma como por la activa comunidad de desarrolladores. Cada módulo se instala independientemente, brindando opciones para evaluar o reforzar una parte del desarrollo del curso.

Cabe resaltar que el uso de Paideia en todos los cursos de la universidad no es obligatorio. Entre el 2011 y 2012 se realizó un proyecto de integración de los datos del Campus Virtual (sistema administrativo de la Universidad) con Paideia, resultando en la creación en simultáneo de todos los cursos de ciclo regular tanto en Campus Virtual como en Paideia. Esta creación de cursos incluye la matrícula de docentes, alumnos, jefes de práctica, así como otros roles con permisos. Al realizarse esta integración, se permitió que todos los cursos tuvieran un espacio en Paideia automáticamente y se

[1] http://www.blackboard.com/about-us/index.aspx
[2] https://www.canvaslms.com/
[3] https://moodle.org/

facilitó que todos los docentes que deseen usar Paideia en sus cursos pudieran hacerlo sin tener que solicitar ningún permiso especial.

Paideia[4] tiene la ventaja de permitir al docente armar su curso con las herramientas (módulos) que él decida. Por ejemplo, cuenta con cuestionarios, tareas para entregar archivos, talleres para promover el trabajo en grupo, wikis para la generación de contenidos por los alumnos, foros de comunicación, agregar recursos de diversos tipos entre otros. Cada docente organiza sus actividades de acuerdo a sus objetivos de aprendizaje, pero muchas veces no cuenta con información sobre que tanto la actividad o recurso fue aprovechada por los alumnos.

Existen diferentes formas en que un docente podría conseguir retroalimentación. Por un lado están las encuestas, las cuales darían una primera mirada sobre la percepción de los estudiantes y la plataforma. Sin embargo, este trabajo se centrará en otro tipo de análisis, enfocado en el área de analítica educativa como la describe Ferguson (2013). Esta área asume la existencia de datos ya guardados que sean legibles por computadoras y que las técnicas a utilizar sirvan para trabajar con grandes cantidades de datos. Para ello, usando como referencia al trabajo del equipo de la Dra. Abdullah (Abdullah, Herawan, Chiroma & Deris, 2014) se decidió usar los registros de logs de la plataforma, los cuales almacenan el comportamiento e interacción real de los alumnos con la plataforma. Los logs presentan la ventaja de expresar hechos reales y no estar afectados por interpretación.

La plataforma cuenta con un sistema de logs que indica a detalle las interacciones que ha tenido cada usuario con cada parte del curso. A pesar de ello, su visualización es engorrosa y no permite la agrupación de elementos bajo ninguna característica, a nivel de curso.

Así mismo, no se cuenta con estadísticas automáticas a nivel de plataforma. Actualmente sólo se calcula manualmente el número de cursos utilizados[5]. Es además importante notar que no se hace una revisión de ciclo en ciclo de cómo van variando los patrones de participación de los estudiantes tanto a nivel de curso como a nivel de plataforma. Cada curso existe, de cierto modo, como su propio universo, y no se busca identificar coincidencias entre ellos.

En el presente trabajo, se propone la utilización de herramientas de minería de datos y aprendizaje de máquinas orientado al estudio de los registros de Paideia con el objetivo de caracterizar los cursos de acuerdo a patrones de comportamiento de los alumnos. El objetivo de este análisis es poder brindar una retroalimentación a los

[4] https://paideia.pucp.edu.pe/
[5] Se define como “curso utilizado” a aquel curso en el que se ha registrado por lo menos un archivo o se ha creado una actividad a lo largo del ciclo.

docentes acerca de cómo utilizan los alumnos el material que generan a fin de mejorarlo.

1.2. Contexto del Problema

Paideia es una plataforma educativa basada en Moodle (moodle.org). Actualmente se utiliza en cursos de pregrado, cursos virtuales y programas/diplomas especiales.

Moodle es un LMS (*Learning Management System*: Sistema de Administración del Aprendizaje) de código abierto que se utiliza en diversas instituciones académicas alrededor del mundo. Cuenta con múltiples módulos desarrollados tanto por los que mantienen el núcleo de la aplicación así como desarrollados por la comunidad. Cada módulo se instala independientemente y cuenta con opciones específicas.

1.2.1. Paideia en la PUCP

Desde la fundación de la Dirección de Informática Académica (Nov. 2005), se hicieron esfuerzos para el manejo de una plataforma LMS independiente del Campus Virtual. Ya en el 2008, se contaba con Diplomados y Cursos Virtuales en la plataforma. Sin embargo, su uso no se hacía aún masivo.

En el 2010, se integró el acceso a Paideia con el de Campus Virtual. Entre el 2011 y 2012 se realizó un proyecto de integración de los datos de Campus Virtual con Paideia, resultando en la creación simultánea de todos los cursos del ciclo regular tanto en Campus Virtual como en Paideia. Estos cursos incluyen la matrícula de docentes, alumnos, jefes de práctica y otros roles.

En el 2013, Paideia pasó a la rama v2 de Moodle usando la versión 2.3, siendo necesario un cambio de servidor y de la base de datos. Es con estos datos con los que se plantea trabajar en este proyecto de tesis, los cuales incluyen datos de los ciclos académicos (periodos de Marzo-Julio, Agosto-Diciembre, Enero-Marzo) del 2013-0 a 2014-2. Se considerarán sólo los cursos que hayan tenido al menos un recurso o actividad creada durante el ciclo.

1.2.2. Módulos en Paideia

Los módulos de aprendizaje de Moodle son componentes que se integran a cada curso de acuerdo a las necesidades de cada docente. Algunos de ellos vienen instalados por defecto y otros se instalan desde el repositorio de módulos de Moodle. Los módulos, a su vez, se dividen en recursos y actividades, siendo éstos últimos los que más interacciones requieren. En Paideia, se encuentran activos y disponibles sólo aquellos módulos que hayan sido previamente validados por la Dirección de

Informática Académica, a fin de que su comportamiento no genere errores tanto para los docentes, al configurarlo, como a los alumnos, al utilizarlo. Siguiendo esta premisa, los módulos de actividad que están disponibles son mostrados en la Tabla 1.

Módulo de actividades	**Identificador en el registro de log**	**Descripción**
Base de Datos	*Data*	Permite la creación de un repositorio mantenible por el docente y los alumnos.
BigBlueButton	*bigbluebuttonbn*	Sirve para generar aulas en línea (tiempo real) para una hora y fecha determinada.
Chat	*Chat*	Genera un espacio de conversación en tiempo real para docentes y alumnos.
Consulta	*Choice*	Sirve para que el docente plantee una única pregunta con opciones.
Cuestionario	*Quiz*	Es el módulo más complejo, ya que incluye la lógica de administración de preguntas y respuestas, así como de feedback.
Encuesta (1) / Encuesta (2)	*questionnaire / feedback*	Son dos módulos que permiten la generación de encuestas de opinión e incluyen plantillas distintas.
Foro	*Fórum*	Genera espacios de comunicación asíncrona, de acuerdo a temas planteados por los docentes. Existen varios tipos de foros, en el que además se puede agrupar a los alumnos.
Glosario	*Glossary*	Sirve para administrar un diccionario de términos en el contexto del curso.
HotPot	*Hotpot*	Permite agregar un recurso interactivo generado por la herramienta HotPotatoes.
Juegos	*Game*	Permite agregar un juego de los tipos: Ahorcado, Crucigrama, Sopa de letras, Millonario, Sudoku, Serpientes y escaleras, Imagen oculta y Libro con preguntas.
Paquete SCORM	*Scorm*	Sirve para importar un paquete que sigue el estándar SCORM (Sharable Content Object Reference Model[6]).
Tarea	*assign / assignment*	Este módulo brinda espacios de entrega de archivos a los alumnos. Existen varios tipos de tareas, de acuerdo a lo que desee lograr el docente.
Wiki	*Wiki*	Genera un espacio colaborativo de creación de páginas web.

Tabla 1: Módulos de actividades de la plataforma PAIDEIA

[6] Conjunto de estándares y especificaciones que permite crear objetos pedagógicos estructurados.

Los módulos de recursos sirven para compartir archivos y enlaces. Los tipos de recursos son: archivo (*resource*), carpeta (*folder*), etiqueta (una imagen o texto que aparece en la página central del curso - *label*), una página web (creada por el docente - *page*) o una URL (*url*).

Aunque no existen propiamente como módulos, algunas funcionalidades como el calendario (*calendar*), agrupar alumnos (*group*), revisar información de los usuarios (*user*) o administrar el curso globalmente (*course*), también aparecen en el log como si fuesen módulos.

1.3. Registro de Log

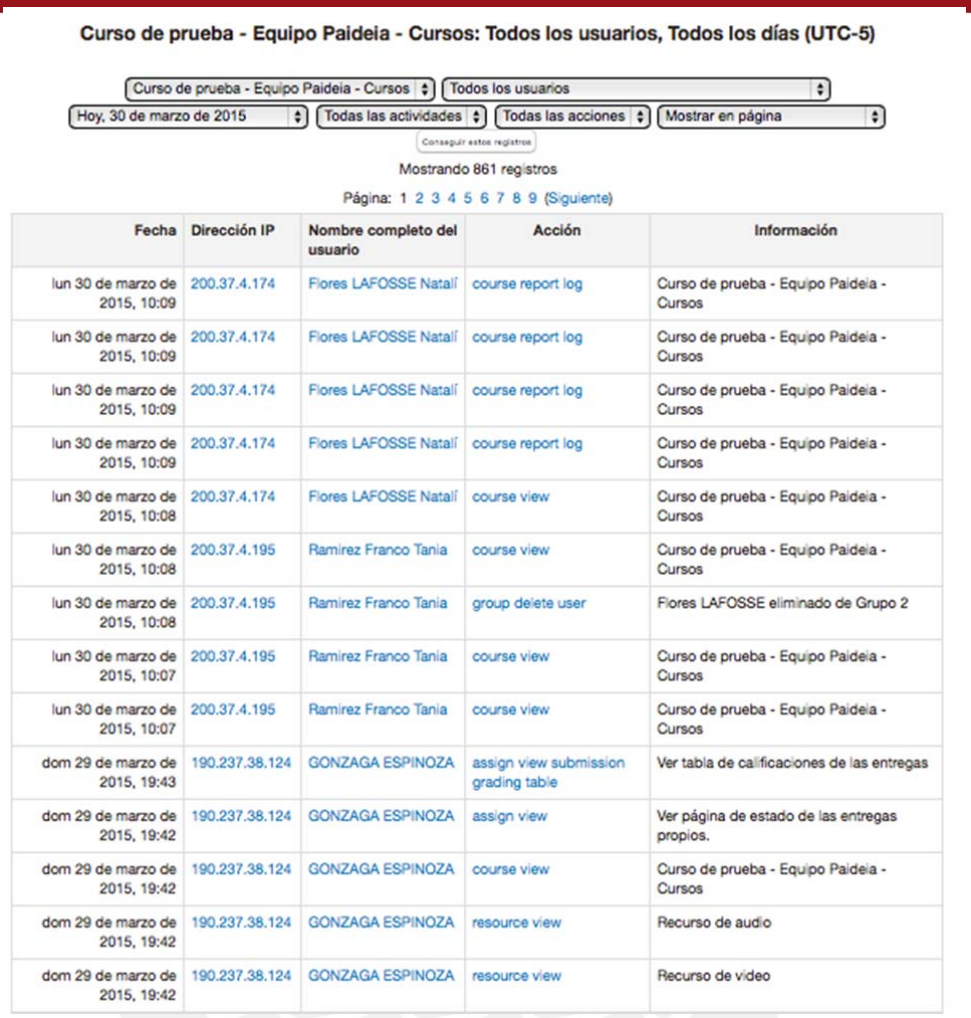
Para el desarrollo del análisis de patrones se tendrá como base la tabla de registro de logs de Paideia, por tanto, es necesario analizar su comportamiento. Por cada visita o transacción que realiza el usuario en la plataforma, Moodle guarda un registro en la tabla indicada. Los registros de esta tabla pueden ser accedidos por los docentes y administradores de la plataforma a nivel de curso y/o plataforma de acuerdo a los permisos con que ellos cuenten. Se puede ver un ejemplo de los registros de Paideia en la Figura 1.

El esquema general de un registro de esta tabla, se visualiza en la Figura 2. En ella se muestran las relaciones más importantes entre un registro y tres tablas maestras: curso (*mdl_course*), usuario (*mdl_user*) y módulo (*mdl_modules*). Estas tres tablas a su vez brindan información respecto a la acción realizada, lo cuál será discutido posteriormente.

Se visualizan como ejemplo, en la Figura 3, los últimos 5 registros de la tabla mdl_log, tomados en Abril del 2015, donde se destacan los usuarios, sus acciones, sus cursos, sus IPs, entre otros campos.

En el campo de acción (*action*), existen acciones comunes como como las vistas (*view*) o de edición (*edit*) y otras específicas de módulos como puede ser el manejo del registro de notas (*grade*), visualización de informes (*report*) y muchas otras.

Es notable además que, gran parte del detalle de los logs se relacionen con datos que se encuentran en otras tablas, así como información específica de cada módulo. En el caso de los cursos, por ejemplo, se encuentran categorizados por facultad y por ciclo por medio del dato *category*, lo cual supondrá una expansión de los datos de log.



Curso de prueba - Equipo Paideia - Cursos: Todos los usuarios, Todos los días (UTC-5)

Curso de prueba - Equipo Paideia - Cursos | Todos los usuarios | Hoy, 30 de marzo de 2015 | Todas las actividades | Todas las acciones | Mostrar en página | Conseguir estos registros

Mostrando 861 registros

Página: 1 2 3 4 5 6 7 8 9 (Siguiente)

Fecha	Dirección IP	Nombre completo del usuario	Acción	Información
lun 30 de marzo de 2015, 10:09	200.37.4.174	Flores LAFOSSE Natalí	course report log	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:09	200.37.4.174	Flores LAFOSSE Natalí	course report log	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:09	200.37.4.174	Flores LAFOSSE Natalí	course report log	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:09	200.37.4.174	Flores LAFOSSE Natalí	course report log	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:08	200.37.4.174	Flores LAFOSSE Natalí	course view	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:08	200.37.4.195	Ramirez Franco Tania	course view	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:08	200.37.4.195	Ramirez Franco Tania	group delete user	Flores LAFOSSE eliminado de Grupo 2
lun 30 de marzo de 2015, 10:07	200.37.4.195	Ramirez Franco Tania	course view	Curso de prueba - Equipo Paideia - Cursos
lun 30 de marzo de 2015, 10:07	200.37.4.195	Ramirez Franco Tania	course view	Curso de prueba - Equipo Paideia - Cursos
dom 29 de marzo de 2015, 19:43	190.237.38.124	GONZAGA ESPINOZA	assign view submission grading table	Ver tabla de calificaciones de las entregas
dom 29 de marzo de 2015, 19:42	190.237.38.124	GONZAGA ESPINOZA	assign view	Ver página de estado de las entregas propios.
dom 29 de marzo de 2015, 19:42	190.237.38.124	GONZAGA ESPINOZA	course view	Curso de prueba - Equipo Paideia - Cursos
dom 29 de marzo de 2015, 19:42	190.237.38.124	GONZAGA ESPINOZA	resource view	Recurso de audio
dom 29 de marzo de 2015, 19:42	190.237.38.124	GONZAGA ESPINOZA	resource view	Recurso de video

Figura 1: Pantalla que muestra los registros de Paideia

Los datos de los usuarios por otro lado, indican el rol de estos en dicho curso, así como ciertas características del usuario. El *timestamp* provee el día y hora en que la acción se realizó, lo cual podría caracterizarse además como días de semana o fin de semana, y horarios de las acciones.

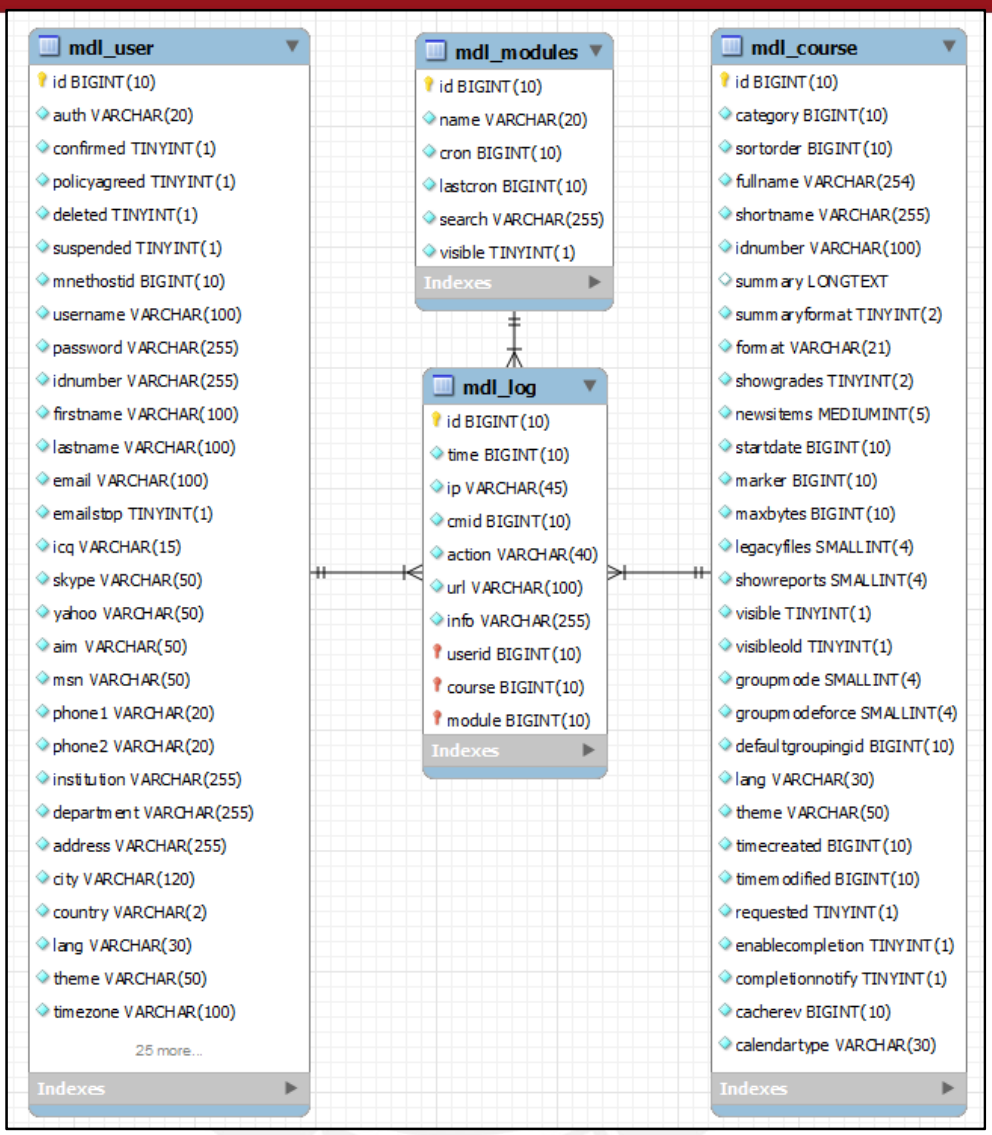


Figura 2: Esquema de base de datos de la tabla mdl_log

id	time	userid	ip	course	module	cmid	action	url	info
8891172	1427744388	2	200.37.4.195	1	user	0	login	view.php?id=2&course=1	2
8891153	1427215740	6143	190.234.30.248	17930	folder	95429	view	view.php?id=95429	4918
8891152	1427215738	20495	190.41.18.79	18558	course	0	view	view.php?id=18558	18558
8891151	1427215737	35558	181.177.248.130	17362	resource	96654	view	view.php?id=96654	42687
8891150	1427215736	5885	200.37.4.28	20003	course	0	view	view.php?id=20003	20003

Figura 3: Los cinco últimos registros de la tabla mdl_log

Capítulo 2. Estado del Arte y Marco Teórico

A fin de encontrar estudios previos de analítica educativa y su uso en sistemas de administración del aprendizaje se realizaron búsquedas en *la base de datos* Scopus usando los términos lms, educational analytics y learning analytics, dentro del título, etiquetas y abstract en Abril del 2015. Con los resultados se obtuvo una lista de 35 estudios, a los cuales se les asignó un puntaje (de -1 a 2), en donde -1 son aquellos que no trabajan el tema planteado (más que todo se centraban en contenidos de cursos); el 0, revisiones de trabajos de terceros; 1, trabajos más aplicativos y 2, aplicaciones de extracción de patrones más avanzadas y relevantes. En las siguientes secciones se revisarán todos los trabajos con puntaje mayor a -1.

2.1 Analítica Educativa y revisiones existentes

El uso de plataformas virtuales en la educación se ha ido masificando desde inicios de siglo, generando gran cantidad de datos sobre el proceso educativo lista a ser explotada. Para analizar dichas cantidades de datos han surgido dos comunidades de investigación (Vishwakarma, 2014): Educational Data Mining (EDM) y Learning Analytics and Knowledge (LAK). Aunque ambas tratan sobre la analítica de datos educativos, se diferencian en que EDM se centra más en el uso de herramientas automáticas y modelamiento generalizado, mientras que LAK está orientado a ayudar a los docentes, buscando resultados para diferentes *stakeholders*.

Otras revisiones de literatura, como la de Romero & Ventura (2010), se centra más en los diferentes resultados que se pueden obtener del análisis de los datos, como es la localización de comportamientos indeseados de los alumnos (plagio), agrupación de estudiantes por performance y recomendaciones para los alumnos.

Castro, Vellido, Nebot & Mugica (2007) realizan una revisión de los distintos algoritmos de aprendizaje supervisado que se utilizan para clasificar datos, de acuerdo a los objetivos que cada estudio desea lograr. Así se tiene, *fuzzy logic (*lógica difusa*)* para predecir el desempeño de los alumnos al resolver ejercicios, redes neuronales para identificar estrategias de navegación, algoritmos evolutivos para identificar métodos de estudio, grafos y árboles para personalizar el contenido del estudiante, reglas asociativas para evaluar actividades, ontologías para encontrar patrones de navegación y algunos sistemas multi-agente para sugerir contenidos. En otros usos se encuentra el clustering con *fuzzy logic* para agrupar por dificultad y EM para agrupar alumnos por comportamiento; regresiones y bayesianos para predecir desempeño y algunos pocos esfuerzos en visualización, que será un punto aparte.

A continuación se clasifica los distintos estudios encontrados de acuerdo al aporte que cada una brinda y que servirán para definir lo que finalmente se realizará en el trabajo. A menos que se indique lo contrario, la plataforma educativa utilizada es basada en Moodle.

2.2 Clustering y clasificación

Entre los estudios que se auto-clasifican como EDM, se encuentran los trabajos de la Universidad de Córdova (Romero, Ventura & García, 2008; Romero, González, Ventura,del Jesus & Herrera, 2009; Romero, Espejo, Zafra, Romero & Ventura, 2013). En ellos, se utilizan algunas herramientas de Minería de datos existentes (Weka) y algunas desarrolladas *in-house* para caracterizar los datos. Sin embargo, dichos análisis se enfocan en clusterizar datos usando distintos modelos (APriori, genéticos, etc.) relacionándola con el resultado de nota final, sin profundizar en los elementos de estas diferencias. El más reciente, del 2013, compara resultados obtenidos con diferentes algoritmos siendo los mejores los de inducción de reglas y *fuzzy logic*, y los de peor resultado: redes neuronales y estadísticos.

El estudio de Bogarín, Romero, Cerezo y Sánchez-Santillán (2014) tiene un objetivo similar, ya que agrupa los datos por medio de resultados; pero luego se enfoca en encontrar patrones dentro de cada clúster. Kotsiantis, Tselios, Filippidi & Komis (2013) y Krpan & Stankov (2012) en sus respectivos estudios, usan clasificación (C4.5) para definir si un alumno aprueba o desaprueba el curso. Sin embargo, Kotsiantis agrega clustering (con k-means) y encuestas de los alumnos, para encontrar reglas de predicción de desempeño.

En el mismo contexto, se encuentra el trabajo de Zhu, Zhang, Wang, Chen y Zeng (2014), en el que se crear clústeres de acuerdo a la navegación y se complementa con el uso de encuestas para medir la emoción. Su clasificación intenta predecir los resultados académicos de los alumnos. Así mismo, se encuentran correlaciones entre las distintas características, generando finalmente una regresión.

Trabajando con la navegación, también se encuentra el trabajo de Blot, Saurel & Rousseaux (2014), el cual utiliza grafos temporales para representar la navegación de los estudiantes, y luego agrupa los nodos de los grafos, tomando en cuenta actividades y recursos para encontrar patrones en ellos. Con ello, se detectan patrones de navegación y de cómo se comportan los alumnos dentro de un curso en específico. Su objetivo se centra en brindar *feedback* a los autores de dicho curso para notar los nodos que mayor visualización tienen. Otro enfoque que utiliza grafos, es el estudio de Martinovi (2012). Sin embargo, él utiliza los grafos para comparar

comportamientos entre alumnos, usando patrones del tipo LCSS (*longest common subsequence*) & TWLCS (*time-warped LCS*), a diferencia del estudio anterior.

A resaltar se encuentran los trabajos realizados por el equipo de la Dra. Abdullah. El primero (Abdullah, Herawan, Chiroma & Deris, 2014), desarrolla una herramienta para determinar patrones secuenciales de comportamiento, a fin de localizar clustering usando los logs de la plataforma educativa. Para ello, realiza un mapeo de las características localizadas en estos registros de logs para luego definir los patrones, eliminar los outliers e informar descriptivamente cómo se comportan los estudiantes.

Una evolución del estudio, es el publicado el año pasado (Abdullah, Alqahtani, Aljabri, Altowirgi & Fallatah, 2015), en el que utilizan dos tipos de datos: estáticos (de cuestionarios) y de navegación (en los logs). Los datos se pre-procesan y se determinan los patrones, clasificándolos usando árboles de decisión (CART, CTA, regresión) de acuerdo a los estilos de aprendizaje definidos por Felder Silverman. Con ello, buscan proporcionar a los docentes mayor información sobre el desempeño de sus alumnos, a fin de mejorar sus materiales.

2.3 Data Warehouse en Educational Analytics

Por otro lado, la forma en que los datos almacenados son trabajados también puede utilizar técnicas de Data Warehouse y esquemas de estrella. El trabajo de Falakmasir, Moaven, Abolhassani & Habibi (2010) busca analizar la participación de los alumnos a través del tiempo. Para ello, define un cubo de cuatro dimensiones: fecha, estudiante, sesión, curso, a partir de un esquema estrella en que la tabla central son los logs de actividades. Con estos datos, se generan gráficos para evaluar cómo varía en el tiempo la participación, determinando que va disminuyendo en su caso. Cabe resaltar que lo aplica en 100 cursos específicos.

Otro datamart fue el creado por Nebić & Mahnič (2010). Este cubo tenía tres dimensiones: actividad, facultad y ciclo, el cual surgió de un esquema estrella basado nuevamente en el log. Su principal objetivo era caracterizar a la población de acuerdo a su participación en el tiempo. Para ello, se generaron reportes relacionando el tiempo de trabajo comparado al tipo de día de la semana, cuánto tiempo en horas se utilizaba para cada tipo de actividad, el cambio de actividades utilizadas a través del tiempo, etc.

2.4 Visualización y análisis de datos asociados a e-learning

Otro reto que se ha encontrado es el de visualizar los datos que se han almacenado en las plataformas. La cantidad de datos que se guarda no suele tener

una forma sencilla de ser visualizada, lo que hace complicado tomar decisiones respecto a ella. Existen algunos esfuerzos, generalmente enfocados en una herramienta en particular. Por ejemplo, Ogashiwa, Hamamoto, Wang, Kariya & Ogawara (2013) plantean una herramienta gráfica que muestra patrones de errores para ayudar a dar retroalimentación a los alumnos que responden cuestionarios.

El trabajo de Florian, Glahn, Drachsler, Specht & Fabregat Gesa (2011) busca modelar parte de los datos, asociándolos a roles, competencias (calculadas) y evaluaciones para visualizar las actividades de los alumnos. Por otro lado, Kapros & Peirce (2014) diseñan una herramienta propia llamada EVADE, para visualizar las actividades del curso (foros, tareas, etc.) realizadas.

Otra forma de visualización, es la definición de rúbricas de acuerdo a las acciones realizadas en la plataforma educativa. El estudio de Dimopoulos, Petropoulou y Retalis (2013) que continúa en otros estudios (Petropoulou, Kasimatis, Dimopoulos, & Retalis, 2014), genera tablas con las diferentes rúbricas definidas por los docentes para evaluar cómo van desempeñándose los alumnos.

Por otro lado, Monk (2005) analiza el tiempo que los alumnos pasan de actividad en actividad, para generar resúmenes y promedios de tiempos. Su visualización se centra en los logs de actividades de la plataforma (en este caso Blackboard), complementado por medio de encuestas, a fin de determinar las preferencias de los alumnos respecto a las diferentes clases de actividades recursos. Es especialmente interesante el procesamiento que realiza para poder luego visualizar los datos.

Existe un estudio realizado por Nagi & Suesawaluk (2008), que además de generar gráficos de visualización, arma reportes comparando el ratio de recursos vistos por los alumnos respecto a la interacción realizada por ellos. Este tipo de estudio se enfoca más al tipo de acción que realiza el alumno, además de su acceso a la actividad, y ya entraría en la temática de LAK.

Finalmente, mencionar el trabajo de Kato & Ishikawa (2013), en la búsqueda de fallas en el aprendizaje, buscando patrones de errores en las respuestas de ciertos cuestionarios, para encontrar donde los rangos de respuestas correctas están bajo el promedio y así detectar los temas a reforzar.

2.5 Metodología KDD: Descubrimiento de Conocimiento a partir de bases de Datos

En principio, uno de los más grandes desafíos de la tecnología actual, es encontrar la manera de utilizar y analizar los grandes contenidos de datos que se obtienen día a día. El abaratamiento de los métodos de almacenamiento, así como la mejora de las tecnologías hace necesario el uso de nuevos métodos y procedimientos

para encontrar información dentro de alto volumen de datos, para así poder tomar decisiones de mejoras o cambios.

Para facilitar el descubrimiento de conocimiento, es decir, de información importante dentro de grande volúmenes de datos, Fayyad, Piatetsky-Shapiro & Smyth (1996) propusieron un marco de trabajo que puede ser adaptado a diferentes tipos de problemáticas y soluciones.

Ellos consideraron una serie de pasos que puede ser resumida en la Figura 4, que serán explicados a continuación.

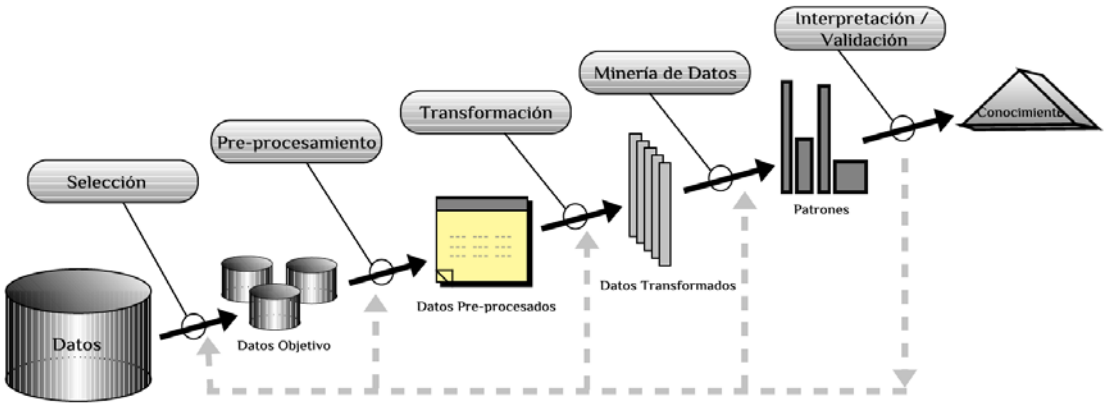


Figura 4: Pasos del proceso de KDD, propuesto por Fayyad, Piatetsky-Shapiro & Smyth (1996)
Traducción propia

Lo primero es determinar el dominio en que se va a trabajar. Luego, se realiza una selección de los datos que se va a utilizar evaluando las secciones de los datos que contengan la mayor información o que presentan algunas características especialmente interesantes.

Una vez seleccionados los datos a utilizar, se realiza el pre-procesamiento. Este paso puede incluir técnicas de limpieza de ruido de los datos, manejo de datos faltantes o retiro de datos incompletos. El siguiente paso es la transformación de los datos, en el cuál se puede reducir la dimensionalidad o proyectar algunas variables de los datos que pueden cambiar la forma en que se muestran.

Con estos datos transformados, se seleccionan los métodos de minería de datos (que se explicarán más adelante) que permitan cumplir con los objetivos. Una vez definido, se empieza con el análisis exploratorio de datos. Éste se refiere a la selección de técnica(s) y algoritmo(s) que se utilizarán para atacar el problema. Es también importante mencionar que se usarán algunas herramientas de visualización para ir descubriendo modelos en los datos.

Después de decidir las técnicas y herramientas, se prosigue con la Minería de Datos en sí, la cual permitirá extraer patrones usando las técnicas seleccionadas

anteriormente. Los patrones descubiertos serán validados, por expertos, a fin de usar el conocimiento que aportan en la toma de decisiones o generación de cambios.

2.5.1 Base de datos a disposición

Los datos con los que trabaja KDD pueden ser obtenidos de diversas formas, sea por sensores o ingresados manualmente. Es también común, el trabajar con logs. Se define a un log como un registro en que se guarda las acciones realizadas en un sistema. De acuerdo a las características propias del sistema, pueden guardar información acerca de quién realizó la acción y los elementos participantes en la acción identificada (elementos afectados, elementos anteriores, momento de la acción, sección en que se encontraba la acción, de dónde se realizó la acción). La mayor parte de sistemas incluyen algún tipo de registro de logs, sea para realizar auditoría, control de calidad o para identificar patrones en las acciones realizadas.

Algunas de las características globales de los logs son el uso de la dimensión del tiempo, así como la del espacio en algunos casos, cuando los datos están asociados a una posición espacial.

2.5.2 Minería de Datos

Se ha definido la Minería de datos como una parte del proceso KDD. Este concepto también se define como como la búsqueda de patrones y correlaciones entre datos de un mismo grupo. Castro *et al.* (2007) declaran que estos métodos son el puente entre la estadística tradicional, reconocimiento de patrones y aprendizaje de máquinas. Todos los métodos de minería de datos buscan generar modelos relativos a los patrones que se encuentran en los datos para poder caracterizarla o predecir resultados en algunos casos.

Los métodos de Minería de Datos pueden agruparse de acuerdo a la forma en que trabajan los datos (Palace, 1996):

2.5.2.1 Aprendizaje Supervisado (Clasificación) y Predicción

Se llama así a los métodos que utilizan las características de los datos para clasificarlos de acuerdo a clases previamente determinadas. Para ello se generan modelos que permiten predecir las clases de acuerdo a las características seleccionadas (Mahmoud, 2008). Algunos de los algoritmos utilizados en esta clase de análisis son: árboles de decisión, modelos bayesianos, clasificación por reglas, Support Vector Machines (SVM) y otros más (Han, 2014).

2.5.2.2 Aprendizaje No Supervisado (Clustering)

El análisis de clustering busca conseguir agrupaciones de los datos de acuerdo a características comunes, como se visualiza en la Figura 5. Se diferencia de la clasificación, porque los datos no cuentan con una etiqueta o clase previa (Han, 2014) y el número de grupos se estima de acuerdo al algoritmo que se elija utilizar. El objetivo del clustering es obtener alta separación (poca similaridad entre clases) y homogeneidad (alta similaridad intra-clases) (Xiong, 2008). Se pueden distinguir diferentes enfoques al problema como: clustering por partición (k-means, k-medoids), clustering por jerarquía (Diana, Birch), por densidad (DenClue, Optics), y otros (Xiong, 2008).

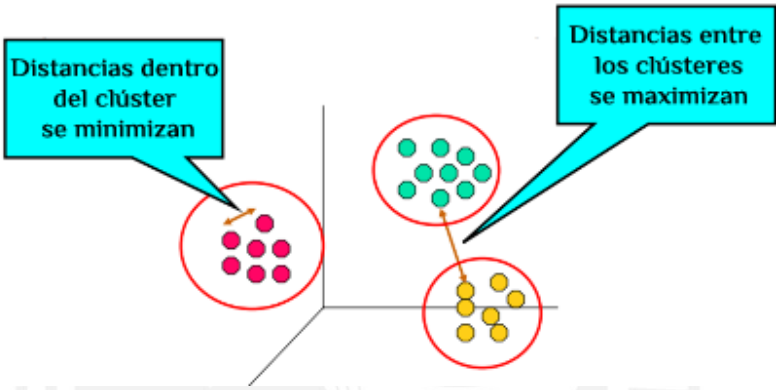


Figura 5: Ejemplo de clústeres (Xiong, 2008) (traducción propia)

2.5.2.3 Reglas de Asociación e Itemsets Frecuentes

Los itemsets frecuentes suponen la lista de elementos que frecuentemente suelen encontrarse juntos. A partir de los itemsets frecuentes, se pueden extraer reglas de asociación, que representan la asociación entre las características, marcando una correlación entre ellas. Se diferencia de las reglas de clasificación y conjuntos frecuentes en que al lado derecho de las reglas puede aparecer cualquier par o pares atributo-valor (Morales & Gonzales, 2013) y que se relacionan a otro par o pares implicando una correlación. Se evalúan de acuerdo a su soporte (probabilidad de contener un grupo de elementos) y confianza (probabilidad condicional de que dado un grupo se contenga otro elemento). Algunos algoritmos que cabe mencionar son Apriori, GenRules, AssoRules, entre otros.

2.5.3.4 Patrones Secuenciales

Los patrones secuenciales obtienen salidas similares a las reglas de asociación, pero con la marcada diferencia de que los elementos tienen un orden de aparición. (Agrawal & Srikant, 1995). Los algoritmos suelen incluir una parte de generación de

candidatos, para luego ir podando de acuerdo a la frecuencia con que aparecen en los datos. (Masseglia, Teisseire & Poncelet, 2005).

2.6 Visualización

Una vez realizada la minería de datos, es necesario que los patrones sean validados por expertos en el dominio de estudio. Para ello, se utilizan técnicas de visualización, al momento de presentar los resultados.

Se llama visualización a la forma en que se presenta gráficamente a los datos (Miccoli, 2013) de manera que puedan ser interpretables por aquellos que toman las decisiones. Es una pieza clave en los análisis exploratorios porque muestra de manera clara la forma que tienen los datos y puede apoyar a localizar patrones antes de realizar mayores análisis. Debe considerarse además que la multidimensionalidad de algunos datos, presenta desafíos en la forma en que dichos datos pueden ser mostrados.

Como se indicó anteriormente, la visualización también puede utilizarse para mostrar los modelos resultantes de la Minería de Datos.

2.6.1 Grafos

Una tendencia actual es la utilización de grafos para representar las relaciones generadas por los modelos de aprendizaje de máquinas (University of Hamburg, 2015). Muchos de los algoritmos han considerado en su propia definición el uso de grafos, como por ejemplo detección de comunidades en grafos, particionamiento de grafos para clustering, kernels de grafos para describir la similaridad de grupos y otros más (University of Hamburg, 2015).

Capítulo 3. Objetivos, Hipótesis y Variables

3.1 Objetivo General

Explotar los datos de un entorno virtual de aprendizaje a fin de caracterizar el comportamiento de los individuos a partir de las acciones realizadas por los usuarios en la plataforma Paideia.

3.2 Objetivos Específicos

1. [O1] Construir un conjunto de datos basado en la tabla de logs que incluya toda la información asociada a la acción realizada, a fin de analizarla.
2. [O2] Utilizando diferentes algoritmos de minería de datos, identificar agrupación de individuos de Paideia por características.
3. [O3] Extraer patrones que permitan identificar correlaciones entre las características de los individuos existentes en Paideia.
4. [O4] Extraer patrones que representen correlaciones temporales entre características asociadas a los logs de Paideia.
5. [O5] Visualización de algunos de los resultados.

3.3 Resultados esperados

Para cada objetivo específico:

- Definir una lista de etapas realizadas para el pre-tratamiento de datos. [O1]
- Tres conjuntos de datos para utilizar con ellos las tres técnicas distintas. [O1]
- Resultados del proceso de clustering. [O2]
- Agrupación de listados de reglas de asociación dependientes a diferentes características. [O3]
- Lista de patrones secuenciales descubiertos [O4]
- Hacer una comparación semántica de los tres métodos antes discutidos. [O5]
- Implementar una herramienta visual que permita presentar los resultados para retroalimentación. [O5]

3.4 Hipótesis

Se puede extraer tres tipos de patrones semánticamente diferentes a partir de los datos asociados a los logs de Paideia.

3.5 Variables

Las variables a utilizar son las características que se pueden encontrar en cada registro de log, las cuales se conocen con el nombre de dimensiones. Se encuentra un recuento de ellas en la Tabla 2.

Nombre del valor	Tipo de dato	Descripción	Valores que puede contener	Tablas Involucradas
Usuario	Nominal	El usuario que realizó la acción	Se manejan los identificadores de usuarios, para no violar la anonimización de los datos.	mdl_user
Rol de Usuario en Curso	Nominal	Tipo de usuario en el curso	Docente, Alumno, Administrador de plataforma, Participante, Participante con permisos, Jefe de Práctica, Jefe de Práctica son permisos	mdl_role_assignments
Curso	Nominal	Curso en el que se realizó la acción	IDs del curso	mdl_course
Ciclo	Ordinal	Ciclo en que se dictó el curso	2013-0, 2013-1, 2013-2, 2014-0, 2014-1, 2014-2	mdl_course_categories
Facultad	Nominal	Facultad a la que pertenece el curso	Id y nombre de la categoría relativa a la facultad	mdl_course_categories
Módulo	Nominal	Contexto de la acción, puede ser actividad, recurso o administrativo	Palabra identificadora del contexto	mdl_modules (y las relativas a su módulo)
Tiempo	Entero	*Timestamp* del log	Microsegundos desde 31/12/1969	-
IP	Intervalo	IP público desde donde se realizó la acción	Considerar que todos los ingresos desde el mismo campus tienen el mismo IP.	-
Acción	Nominal	Acción realizada sobre el contexto identificado	Add, Admin, Assign, Attempt, Choose, Close, Comment, Continue, Create, Delete, Diff, Download, Edit, Enrol, Error, Grade, History, Index, Launch, Lock, Login, Logout, Mailer, Map, Move, New, Preview, Recent, Remove, Report, Restore, Revert, Review, Search, Submit, Subscribe, Startcomplete, Talk, Templates, Unenrol, Unassign, Unlock, Unsubscribe, Update, View. Se incluyen acciones específicas de módulos.	-
Info	Nominal	Más información sobre el módulo involucrado	Puede ser el id del módulo, el perfil de usuario involucrado u otras informaciones adicionales.	Relativas a su módulo

Tabla 2: Dimensiones de los logs

Capítulo 4. Experimentación y Resultados

Para el desarrollo del análisis de patrones se utilizará KDD, siendo la base del análisis la tabla de registro de logs de Paideia. Para ello se inicia con una exploración de los datos, para luego proceder a seleccionarlos y seguir con el procedimiento ya descrito (selección, pre-tratamiento, transformación, minería y evaluación).

4.1 Exploración de datos

Para poder seleccionar el conjunto de datos a utilizar, es necesario explorar las bases de datos con la que se cuenta. Se ha explicado anteriormente las dimensiones con las que una acción cuenta, sin embargo dichos datos se encuentran disgregados en diferentes tablas.

Se empieza con generar tablas de agrupamiento según algunas de las dimensiones identificadas de los *logs*. Se ha utilizado sólo los 8 145 904 registros que corresponden a los ciclos indicados. Los registros que no existen en dichos ciclos corresponden a cursos de pruebas, ejemplos de uso o datos pertenecientes a ciclos incompletos[7], por lo que no se utilizarán en el estudio.

La Tabla 3 resume la forma en que se distribuyen los registros por ciclo, incluyendo el promedio de registros que pertenecen a cada curso y la desviación estándar calculada. Se nota que aunque los promedios no pasan de mil registros por curso, la desviación estándar asciende a cuatro o cinco veces el promedio calculado. Ello indica que existe mucha variabilidad entre los registros que se van a analizar.

La Figura 6 ilustra la información de cursos por ciclo. En ella se muestra una menor cantidad de cursos con actividad (es decir, en los que exista al menos un registro en la tabla de logs) respecto a los ciclos regulares del 2013 al 2014. Sin embargo, esta información se contrasta con la Figura 7, en la que se muestran los registros de actividad por ciclo, en el que se nota un claro incremento de actividad del 2013 al 2014.

Por otro lado, la Tabla 4 lista los dos cursos con mayor cantidad de registros en cada ciclo a evaluar. Estos números indican que los cursos con mayor actividad manejan un número de registros que supera los 70 mil en el caso de ciclos durante el año y 10 mil en el caso de ciclos de verano.

Otra distribución a evaluar son las acciones de los registros. En la Tabla 5 se han agrupado los registros por ciclo de acuerdo a las acciones realizadas, agrupando las 21 acciones que contienen una frecuencia menor al 0.1% del total. Se nota a su vez, que más del 84% de los registros se centra en la acción *view*, la cual tiene una proporción de casi 10 veces más que la siguiente acción, *continue*, en cada ciclo

[7] Ciclos que no habían concluido al momento del estudio

evaluado. Las siguientes acciones tienen distribuciones menos uniformes durante los ciclos, la mayor parte superando los miles de registros.

Ciclo	Cursos	Registros	Promedio de Registros por Curso	Desviación Estándar por Curso
2103-0	132	59246	448.83	1987.97
2013-1	4081	1785188	437.44	4128.28
2013-2	3123	1666853	533.73	2970.97
2014-0	211	65904	312.34	1200.90
2014-1	2990	2244540	750.68	4242.21
2014-2	2862	2324173	812.08	3923.49
Total	**13399**	**8 145 904**	**-**	**-**

Tabla 3: Estadísticos de los registros por ciclo

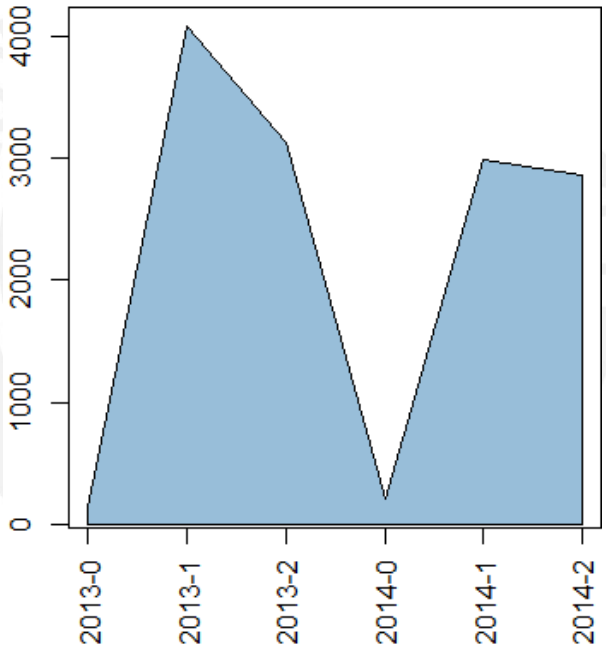


Figura 6: Cursos con actividad por ciclo

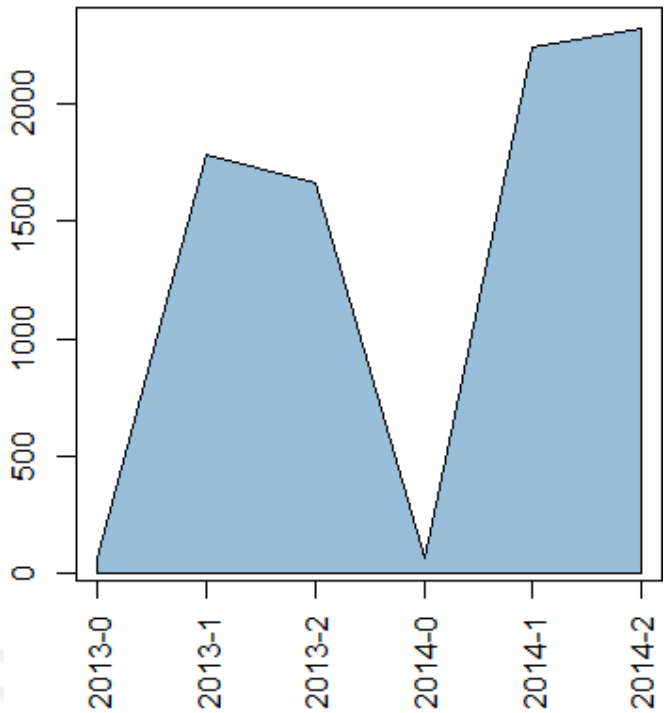


Figura 7: Registros de actividad por ciclo (valores en miles)

Ciclo	Registros	Cursos
	13271	2013-0 APRENDER Y COOPERAR EN REDES SOCIALES (TIC612-C751)
2103-0	12119	2013-0 QUÍMICA 1 (QUI117-0103)
	151350	2013-1 INTRODUCCIÓN A LA COMPUTACIÓN (INF117)
2013-1	150646	2013-1 GESTIÓN PÚBLICA (GEP310)
	96116	2013-2 INTRODUCCIÓN A LA COMPUTACIÓN (INF117)
2013-2	74394	2013-2 TALLER DE HABILIDADES GERENCIALES 1 (GES260)
	11986	2014-0 QUÍMICA 1 (QUI117-0103)
2014-0	7825	2014-0 MOBILE LEARNING Y EDUCACIÓN (TIC614-C751)
	139103	2014-1 INTRODUCCIÓN A LA COMPUTACIÓN (INF117)
2014-1	110120	2014-1 GESTIÓN PÚBLICA (GEP310)
	128187	2014-2 GESTIÓN PÚBLICA (GEP310)
2014-2	80561	2014-2 INTRODUCCIÓN A LA COMPUTACIÓN (INF117)

Tabla 4: Cursos con más acciones por ciclo

Acciones (agrupadas)	2013-0	2013-1	2013-2	2014-0	2014-1	2014-2	Total general	Porcentaje	Porcentaje Acumulado
View	47567	1476695	1451946	56143	1889807	1942543	6864701	84,27%	84,27%
Continue	2393	154928	101580	3964	177222	196683	636770	7,82%	92,09%
Add	1708	24045	25487	1745	41198	33930	128113	1,57%	93,66%
Review	870	41723	13748	164	32279	34233	123017	1,51%	95,17%
Edit	985	15085	14569	463	16441	18932	66475	0,82%	95,99%
Submit	124	9123	13653	303	17177	22940	63320	0,78%	96,77%
Grade	772	10845	7614	404	11143	18050	48828	0,60%	97,36%
Update	1464	9619	8649	391	10100	13588	43811	0,54%	97,90%
Report	693	9871	7271	216	8268	10977	37296	0,46%	98,36%
Delete	584	10408	6321	72	10744	2868	30997	0,38%	98,74%
Attempt	147	7248	4432	100	7687	11282	30896	0,38%	99,12%
Close	145	6755	4015	103	7428	11064	29510	0,36%	99,48%
Talk	1699	2118	142	1264	9425	571	15219	0,19%	99,67%
*otros**	95	6725	7426	572	5621	6512	26951	0,33%	100,00%
Total general	**59246**	**1785188**	**1666853**	**65904**	**2244540**	**2324173**	**8145904**	**100,00%**	**-**

* “otros”: history, comment, preview, map, download, create, launch, search, recent, assign, unsuscribe, suscribe, diff, choose, remove, lock, unlock, admin, index, template, restore, unassigns, enrol, revert, startcomplete, unenrol y move.

Tabla 5: Acciones agrupadas por ciclo, ordenados por frecuencias

La Tabla 6 y la Figura 8 ilustran la agrupación de registros por cada módulo de la plataforma respecto al ciclo. En la Tabla 6 se nota que hay algunas casillas que llegan a valores muy altos, siendo el mayor el módulo de curso (*course*), que agrupa todas las acciones relativas a cursos. En la Figura 8 se muestran, utilizando círculos, las diferencias de densidad de las casillas, a fin de localizar los elementos más frecuentes.

Módulo	2013-0	2013-1	2013-2	2014-0	2014-1	2014-2	Total general
assign	2423	117588	162799	5199	200365	248922	737296
assignment	2	-	-	-	-	-	2
calendar	255	4423	4618	169	6348	8760	24573
chat	1818	3392	343	1462	11343	917	19275
choice	-	1146	337	-	3	132	1618
course	22913	712720	693358	25003	928760	949394	3332148
data	-	-	593	-	348	136	1077
feedback	-	-	-	-	-	8	8
folder	5422	130676	126079	4946	161301	148158	576582
forum	10282	65341	65853	4627	75640	76438	298181
game	-	-	-	-	-	32	32
glossary	-	127	132	-	668	273	1200
group	167	10294	8589	7	20429	3862	43348
hotpot	-	139	14	-	20	45	218
label	364	8280	6553	263	7694	6762	29916
page	512	17855	15006	470	21096	22415	77354
questionnaire	-	1665	1412		2659	2310	8046
quiz	6780	295799	171285	5311	294918	351783	1125876
resource	5286	304068	301166	10893	401618	405073	1428104
scorm	-	1040	1659	-	1435	1070	5204
url	1313	31130	22953	728	31518	25940	113582
user	1708	55925	60242	4638	64487	55184	242184
wiki	1	23580	23862	2188	13890	16559	80080
Total general	**59246**	**1785188**	**1666853**	**65904**	**2244540**	**2324173**	**8145904**

Tabla 6: Registros de módulos agrupados por ciclo

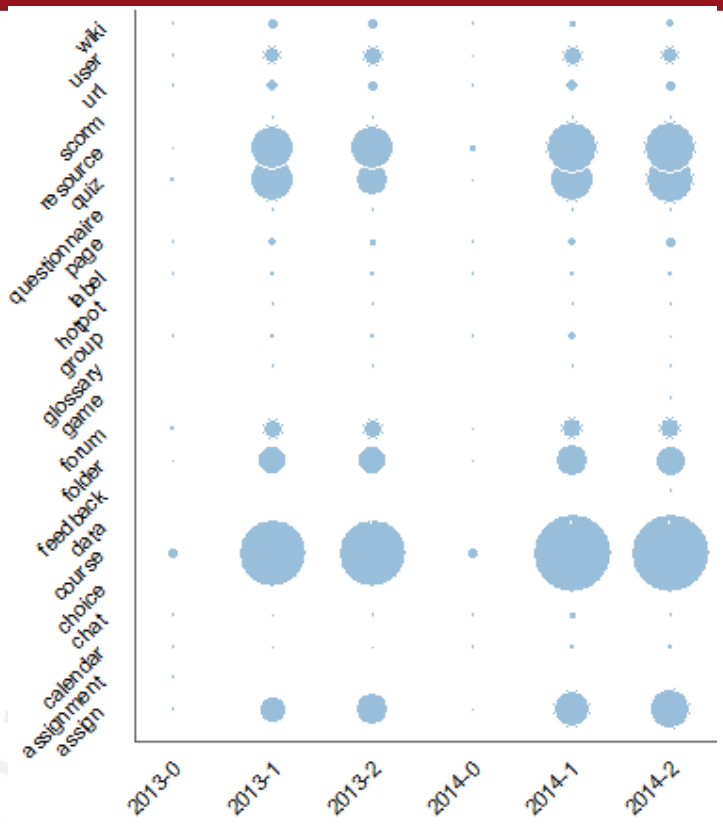


Figura 8: Registros agrupados por modulo para cada ciclo

Además, es interesante mostrar la distribución de los registros de acuerdo a la facultad, que como se indicó anteriormente, se relaciona con la categoría “padre” del curso al que pertenece. En la Tabla 7 se muestra cómo estos registros se distribuyen en las 19 facultades existentes hasta el 2014-2. Las facultades que más actividad presentan son Gestión y Alta Dirección, seguida de las dos facultades de primeros ciclos de pre-grado: Estudios Generales Letras y Estudios Generales Letras. La Figura 9, además, permite visualizar comparativamente la distribución de los registros a modo de áreas de los círculos.

Facultades	2013-0	2013-1	2013-2	2014-0	2014-1	2014-2	Total general
ADMINISTRACION Y CONTABILIDAD	9	745	13381	19	7741	28900	50795
ARQUITECTURA Y URBANISMO	1817	15909	13952	1783	10798	18614	62873
ARTE	10	9384	5418	62	13220	7671	35765
ARTES ESCENICAS	0	5137	2655	15	14193	9372	31372
CIENCIAS E INGENIERÍA	39	90711	210666	6873	155635	270570	734494
CIENCIAS SOCIALES	0	21601	12844	10	23883	18362	76700
CIENCIAS Y ARTES DE LA COMUN.	1	15718	16607	2062	26377	27877	88642
CONSORCIO DE UNIVERSIDADES	0	66	31		16	21	134
DERECHO	31	139467	133255	226	171891	150774	595644
EDUCACION	185	86472	71597	173	86661	76936	322024
ESCUELA DE DANZA CONTEMPORÁNEA	5	7	0	0	0	0	12
ESCUELA DE ESTUDIOS ESPECIALES	0	52	88	0	44	19	203
ESCUELA DE MÚSICA	0	638	0	0	0	0	638
ESCUELA DE POSGRADO	19856	83919	64492	13684	144913	224314	551178
ESCUELA DE TEATRO	0	2	0	0	0	0	2
ESTUDIOS GENERALES CIENCIAS	13331	353486	307495	22135	451608	419839	1567894
ESTUDIOS GENERALES LETRAS	1120	332793	311113	6309	482774	362749	1496858
GESTIÓN Y ALTA DIRECCIÓN	22819	526480	403676	11958	563085	612437	2140455
LETRAS Y CIENCIAS HUMANAS	23	102601	99583	595	91701	95718	390221
Total general	**59246**	**1785188**	**1666853**	**65904**	**2244540**	**2324173**	**8145904**

Tabla 7: Registros de facultades agrupadas por ciclo

Figura 9: Registros agrupados por modulo para cada ciclo

4.2 Selección y pre-tratamiento de Datos

En esta sección, se explicará el criterio utilizado para seleccionar los datos con los que se trabajará, así como las transformaciones necesarias para explotar los datos a fin de cumplir con los objetivos del estudio.

Dada la cantidad de registros con los que se cuenta (superando los 8 millones), se decidió seleccionar un bloque de registros donde aplicar las técnicas de minería de datos sin que esto supusiera un costo computacional demasiado alto. Para ello, el bloque seleccionado debía ser significativo, dado que el objetivo es explorar cómo se comporta la extracción de patrones en los datos.

El primer filtro fue retirar los registros pertenecientes a usuarios externos al curso, como son los administradores o visitantes casuales. Esta condición se basa en que el uso de los administradores al momento de configurar o restaurar los cursos, genera múltiples interacciones que no son relativas al contenido del curso en sí. Así mismo, se filtraron los ingresos a la ruta del curso por personas externas, aunque no

hayan finalmente accedido al contenido. Para ello, se retiraron todos los registros en los cuales el usuario de la acción no tenía un rol asignado a dicho contexto.

Posteriormente se procedió a convertir los campos de “module” y “action” en múltiples campos de acuerdo a su tipo, a fin de obtener mediciones cuantitativas de cada valor. Aquellos valores del campo “action” que tenían poca frecuencia de clics, fueron acumulados en un campo “otro”. A esta transformación de datos, le siguió una agrupación de los datos por semanas y usuarios, para así armar una primera estructura de datos que incluyen una dinámica temporal. Hay que considerar que, como los datos se graban continuamente, la granularidad temporal mínima considerada es el *segundo*. La operación de agrupación que se utilizó fue la acumulación, de manera tal, que los datos representen el total de clics de cada tipo de módulo y acción en cada campo creado.

Después de la etapa anterior, es necesario realizar la selección de datos que finalmente se utilizará para realizar las experimentaciones. Para elegir un grupo de datos de prueba con suficiente variabilidad pero de un tamaño manejable, se designó un protocolo de selección de datos. Se tomaron 10 cursos que cumplieran los siguientes requisitos:

- Debía agrupar más de 99 logs, para ser un curso con actividad no sólo del docente.
- No debía agrupar más de 450 logs, para no elegir cursos con una actividad demasiado alta.
- Debía ser uno de los dos cursos de más logs en un ciclo regular (de menos de 450)
- Debía ser el curso de más logs en un ciclo de verano. (de menos de 450)

Esta selección resultó en un conjunto de datos de 4259 filas que serán las que se utilizarán para realizar la experimentación con los diferentes algoritmos de extracción de patrones. Para cada uno de los algoritmos se requerirán otras transformaciones que se detallarán a continuación.

4.3 Transformación y minería de datos

Para el análisis, se generarán modelos de clústeres, se obtendrán reglas de asociación y patrones secuenciales que descubran correlaciones entre las diversas características de comportamiento de los usuarios de la plataforma. Para la experimentación se utilizó tanto la herramienta Weka (Hall *et al*, 2009) como la librería Java SPMF de Fournier-Viger, Gomariz, Gueniche, Soltani, Wu y Tseng (2014).

4.3.1 Clusterización

Para realizar la extracción de patrones de clústeres, se requirió usar los datos agrupados por usuario, utilizando la acumulación como operación que agrupe todos

los clics según módulo accedido y acción realizada. Esto supuso un conjunto 492 registros, cada uno para un usuario y cada uno sólo para un único curso.

Como no se contaba con un número de clústeres predeterminado para trabajar, se decidió utilizar el Algoritmo EM (Dempster, Laird & Rubin, 1977) a fin de que éste calculase el número de centroides de los datos por sí mismo. El resultado utilizando Weka fueron 3 clústeres. El resumen de la salida se muestra en la Tabla 8.

Tiempo de construcción del modelo	5.49 segundos	
Instancias de clústeres	0	2 (0%)
	1	15 (3%
	2	475 (97%)
Log-Verisimilitud (función soporte)	129.05297	

Tabla 8: Resultados de clúster usando Algoritmo EM

Tomando en consideración que los elementos que se encuentran más frecuentemente son las acciones de “ver” y los módulos “recurso” y “curso” (como se mostró en la sección de exploración de los datos) es interesante visualizar los clústeres utilizando dichos campos para mostrar la dependencia del modelo hacia dichos valores. Las Figuras 10 y 11 muestran estas relaciones.

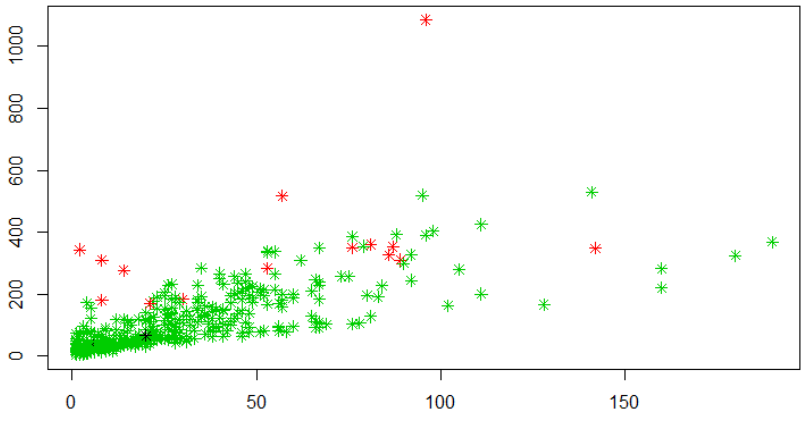


Figura 10: Actividad “ver” vs. módulo “recurso”, clúster por EM

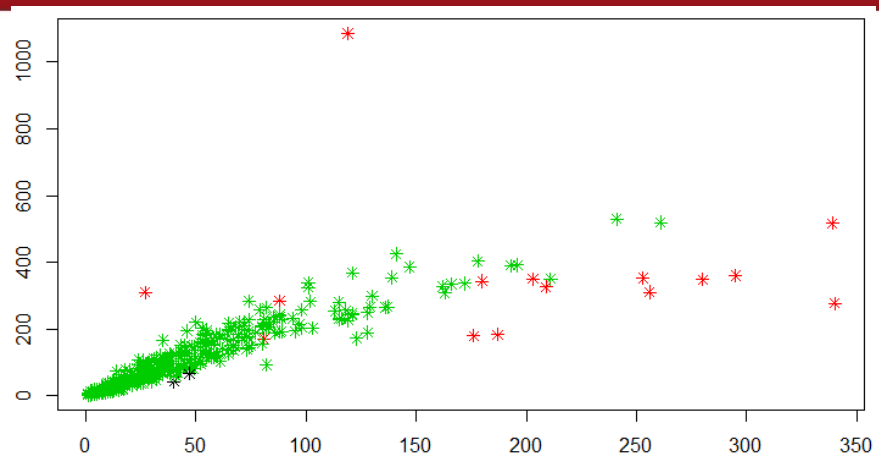


Figura 11: Actividad “ver” vs. módulo “curso”, clúster por EM

Por otro lado, al revisar las diferentes relaciones de otros campos, se encuentran algunas relaciones interesantes. Por ejemplo, en la Figura 12 se muestra la relación de la actividad “agregar” con el módulo “recurso”. Se ve que la distinción de clústeres separa los elementos casi perfectamente. Algo similar ocurre al enfrentar “agregar” con “foro”, dando la impresión que los clústeres usando este modelo dependen en cierta medida de las acciones menos frecuentes, como se muestra en la Figura 13.

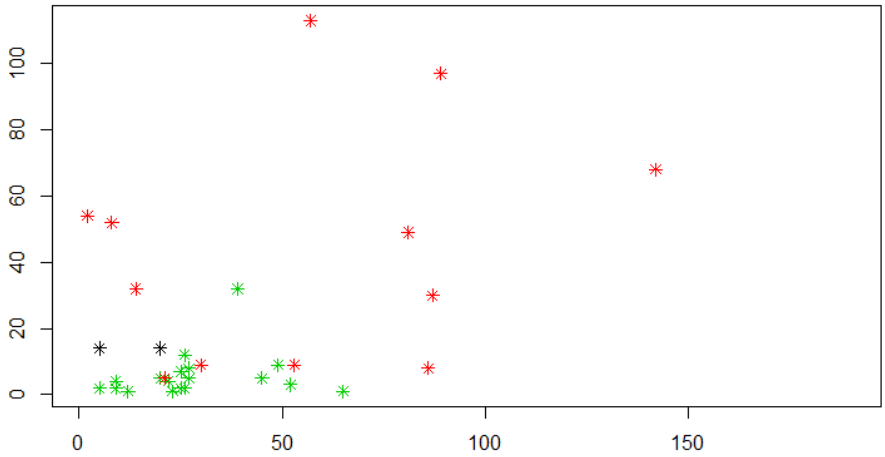


Figura 12: Actividad “agregar” vs. módulo “recurso”, clúster por EM

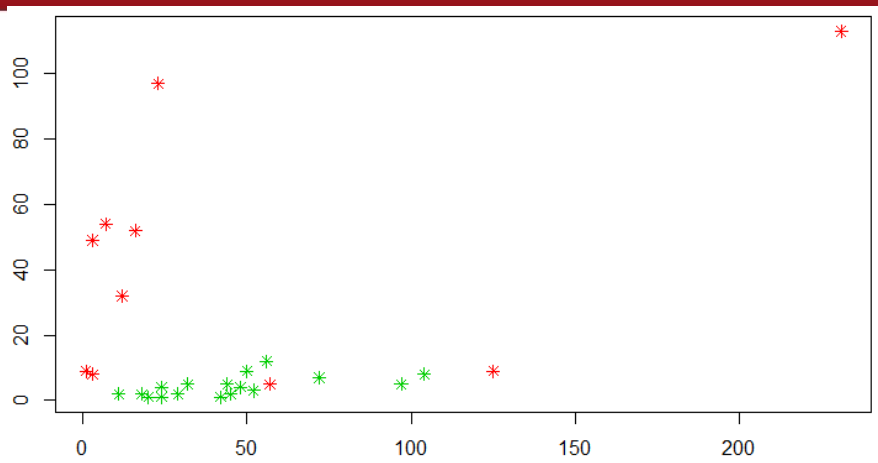


Figura 13: Actividad “agregar” vs. módulo “foro”, clúster por EM

Usando el conocimiento de la existencia de tres clústeres, se procede a realizar una nueva experimentación utilizando otro algoritmo de clusterización, en este caso, Xmeans (Pelleg & Moore, 2000), una extensión al conocido KMeans (MacQueen, 1967) con mejoras en su eficiencia, con la implementación de Weka. Los resultados de esta experimentación se encuentran en la Tabla 9.

Tiempo de construcción del modelo	0.03 segundos	
Instancias de clústeres	0	135 (27%)
	1	319 (65%)
	2	38 (8%)
Distorsión[8]	99.839425	
Valor BIC[9]	12359.107013	

Tabla 9: Resultados de clúster usando Algoritmo Xmeans

En estos resultados, se nota que los clústeres no están tan desbalanceados como los anteriores. Este resultado no sorprende, dado que en cuanto el algoritmo EM trabaja buscando puntos de inflexión maximizando la probabilidad de cada punto, el algoritmo Xmeans utiliza más bien las distancias euclídeas de los puntos para calcular centroides y asume cierta densidad alrededor de ellos. Esto resulta en que Xmeans (considerado una clusterización dura) busque separar distintivamente las poblaciones (Kearns, Mansour & Ng, 1998) a diferencia del algoritmo EM. Para visualizar estas diferencias, se utilizará gráficas similares a las de la clusterización por algoritmo EM.

[8] Suma de las distancias cuadradas desde los centroides
[9] Criterio de Información Bayesiano: Medida la selección de modelos. Da una penalidad ante el *overfitting.* El modelo es considerado mejor mientras menor sea (comparativamente a otro modelo).

Las Figuras 14 y 15 muestran las relaciones de la actividad “ver” con los módulos “recurso” y “curso” respectivamente.

Por otro lado, con este modelo, las relaciones que se vieron con los otros campos ya no se hacen evidentes. La Figura 16, muestra que la separación al mostrar la acción “agregar” y el módulo “recurso”, se diluye.

Hay que considerar que el Algoritmo EM trabaja los datos como un todo, por lo que no sorprende que los campos menos comunes tengan más influencia en su modelo que en Xmeans.

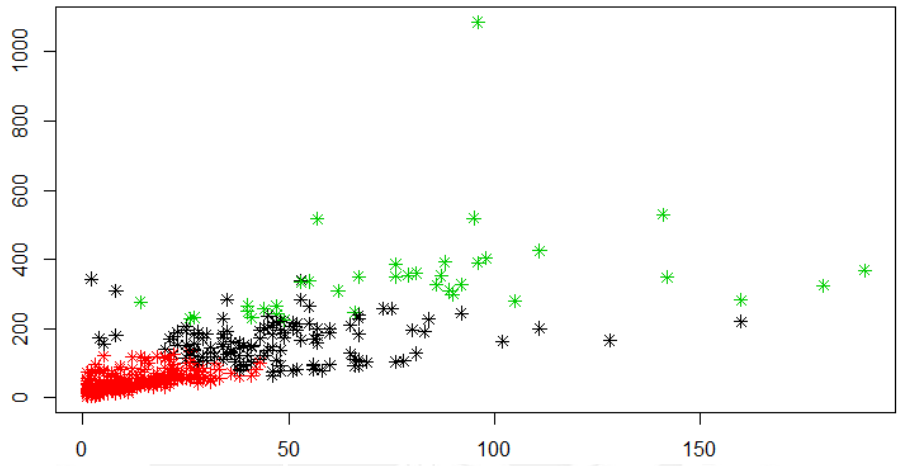


Figura 14: Actividad “ver” vs. módulo “recurso”, clúster por Xmeans

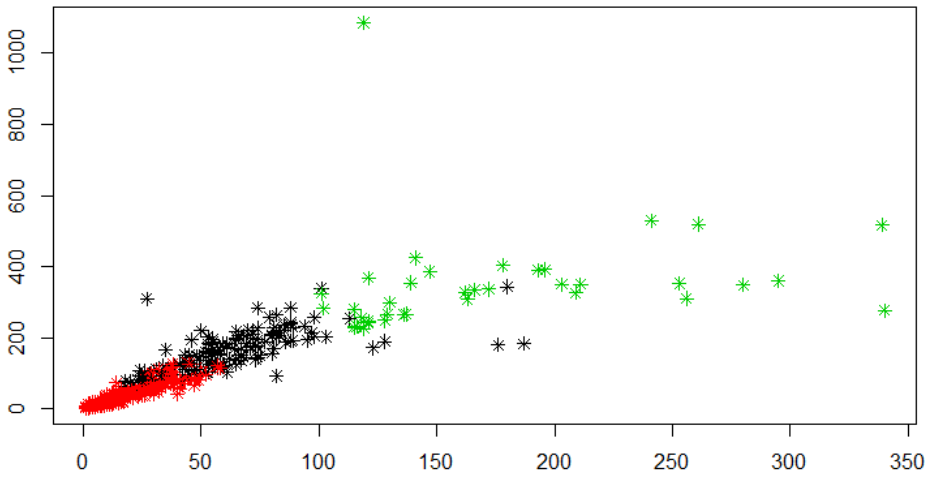


Figura 15: Actividad “ver” Vs. módulo “curso”, clúster por Xmeans

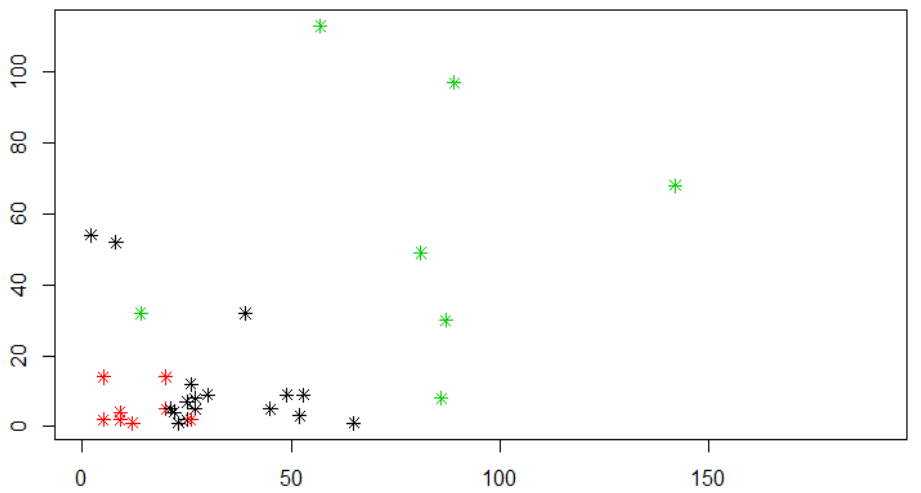


Figura 16: Actividad “agregar” vs. módulo “recurso”, clúster por Xmeans

4.3.2 Reglas de Asociación e Itemsets Frecuentes

Para la extracción de itemsets frecuentes se hizo uso de la librería SPMF con el algoritmo Apriori (Agrawal & Srikant, 1994) y para reglas de asociación (las cuales implican causalidad) la herramienta Weka, con el mismo algoritmo. Ambos resultados tuvieron semejanzas, sin embargo, los obtenidos por Weka expresaban con mayor detalle las relaciones entre elementos al caracterizar mejor la relación de ellos, dándole precedencia a uno sobre el otro. El único parámetro que variará en todas las experimentaciones será el soporte mínimo relativo.

Las reglas de asociación e itemsets frecuentes supusieron una transformación de los datos similar a la utilizada para la clusterización. Nuevamente, se eliminó la dimensión del tiempo de los datos, agrupando por medio de sumas todos los registros semanales correspondientes a un mismo usuario. A este conjunto, se le aplicó un tratamiento más.

Al usar la librería SPMF se requiere la discretización de los datos con los que se cuenta. Para ello, se decidió discretizar utilizando rangos iguales, es decir, separar cada campo en un máximo de seis grupos con rangos de igual longitud, sin importar la cantidad de elementos que perteneciera a cada grupo. Se usó seis rangos, considerando que el mayor rango de los campos tenía como máximo, un valor de 1000. Por otro lado, el menor valor máximo de los rangos era 5. Cada par de campo y rango de valor que existiese (se eliminaron los vacíos) fue etiquetado manera única. Los primeros 10 elementos se muestran en la Tabla 10.

Se realizaron experimentaciones con estos datos, colocando diferentes soportes mínimos (es decir, el mínimo de instancias que cumple la regla) de acuerdo a la especificación de la librería. La librería utiliza un soporte mínimo relativo, en este caso, un porcentaje. Los resultados de cada experimentación se encuentran en la Tabla 11.

Para las experimentaciones se utilizaron soportes cuyo valor variaba entre 0.1 a 0.9, a fin de analizar la variación de los resultados.

Campo	Valores	Etiqueta
actview	<200 & >=1	1
actview	<400 & >=200	2
actview	<600 & >=400	3
actview	>=1000	4
actadd	<22 & >=1	5
actadd	<44 & >=22	6
actadd	<66 & >=44	7
actadd	<88 & >=66	8
actadd	<110 & >=88	9
actadd	>=110	10

Tabla 10: Primeros 10 elementos etiquetados

Soporte mínimo relativo	Tiempo (ms)	Conjuntos	Memoria (Mb)
0.9	0	0	0
0.8	11	3	2.6
0.7	12	3	2.6
0.6	12	3	2.6
0.5	14	7	2.6
0.4	13	7	2.6
0.3	15	13	2.6
0.2	17	32	2.6
0.1	22	87	2.6

Tabla 11: Resultados cuantitativos de experimentación de reglas de asociación (SPMF)

En las Figuras 17 y 18 se muestra el impacto del soporte mínimo sobre el número de conjuntos encontrados y el tiempo de ejecución. No se graficó la memoria ya que se mantuvo constante. Esto se debe a que, el mismo modelo se utiliza para cada una de las experimentaciones, siendo el soporte mínimo lo que descarta a los candidatos. Además, el cambio de conjuntos entre 0.2 a 0.8 tiene una variación mínima. A su vez, a partir del soporte 0.3, el tiempo también se mantiene casi constante, dado que los grupos candidatos son casi los mismos de los cuales se debe elegir.

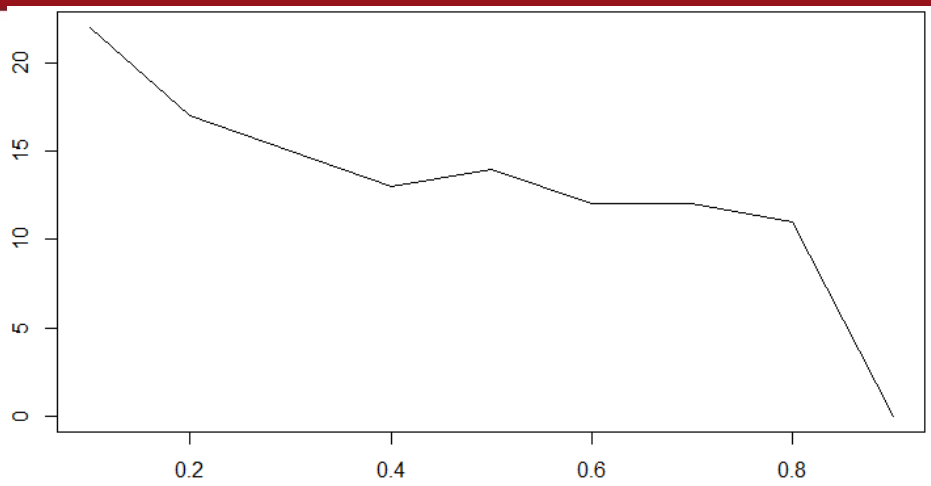


Figura 17: Soporte mínimo relativo vs. número de conjuntos encontrados

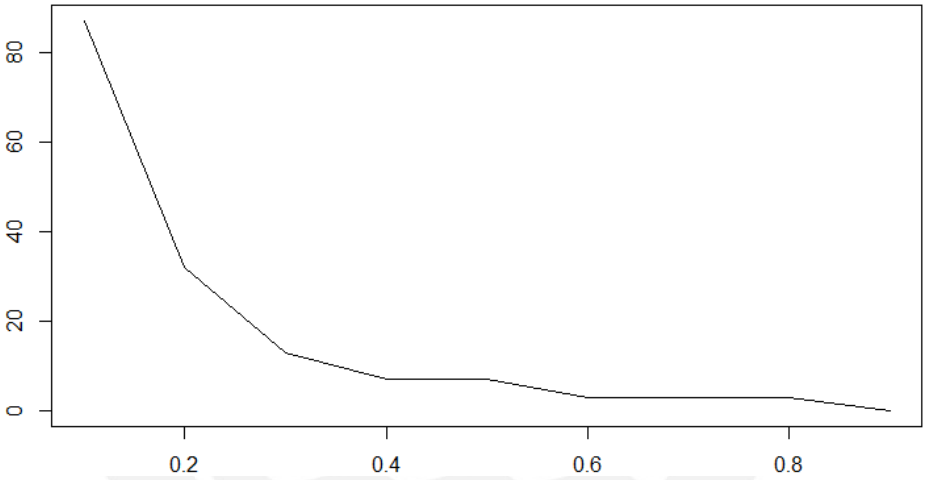


Figura 18: Soporte mínimo relativo vs. tiempo en milisegundos

En la **Tabla 12** se muestran ejemplos de los conjuntos encontrados usando un soporte mínimo de 0.4.

Ítems encontrados juntos frecuentemente	Soporte Absoluto
1=<actview<200	421
1=<modcourse<68	404
1=<modresource<38	289
1=<actview<200, 1=<modcourse<68	398
1=<actview<200, 1=<modresource<38	281
1=<modcourse<68, 1=<modresource<38	268
1=<actview<200, 1=<modcourse<68, 1=<modresource<38	267

Tabla 12: Conjunto de ítems frecuentes usando SPMF, minsup=0.4

Cabe resaltar que SPMF lista como si fueran itemsets, a grupos que sólo contienen un elemento. Es además necesario notar que la salida del SPMF no establece una consecuencia de uno sobre otro, sino que muestra los elementos del conjunto sin distinción, como se espera de los conjuntos frecuentes.

En el caso de Weka se experimentó con soportes mínimos desde 0.1 a 0.9, pero se mantuvo la confianza (proporción de elementos que cumpliendo la premisa, aseguran la consecuencia) en un 0.9. Este parámetro no está disponible en el SPMF, por lo cual puede ser causa de discrepancias en los resultados, así como la necesidad de al menos incluir dos elementos si se busca una causalidad, en Weka. El resumen de las experimentaciones se encuentra en la Tabla 13.

Soporte mínimo relativo	Mejores reglas[10]	Ciclos
0.9	0	0
0.8	2	4
0.7	2	6
0.6	2	8
0.5	7	10
0.4	7	12
0.3	11	14
0.2	28	16
0.1	72	18

Tabla 13: Resultados cuantitativos de experimentación de reglas de asociación (WEKA)

Las Figuras 19 y 20 muestran la relación de los soportes mínimos con el número de reglas y la cantidad de ciclos de procesamiento. Se muestra que en Weka el número de reglas decrece de forma más pronunciada en comparación a los experimentos con SPMF. Por otro lado, el número de ciclos (iteraciones realizadas para seleccionar las reglas) varía de forma escalonada.

Se pueden ver ejemplos de los resultados de la experimentación con soporte mínimo de 0.4 en la Tabla 14. Se utilizan las mismas etiquetas que en la Tabla 12.

Las reglas con soportes mayores a 0.4 tanto en Weka como en SPMF se centran en combinaciones de los siguientes elementos: 1=<modcourse<68, 1=<modresource<38 y 1=<actview<200. Al igual que en los patrones secuenciales, es la actividad de “ver” unida a los módulos de “curso” y “recurso.”

[10] Reglas que satisfacen la confianza y soporte mínimo

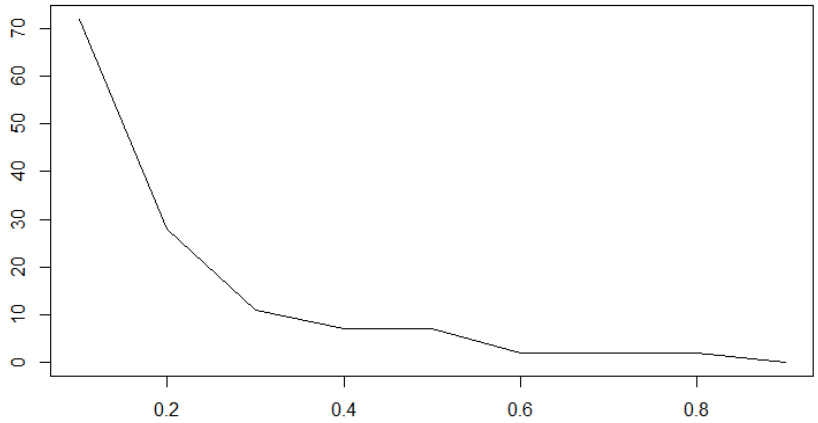


Figura 19: Soporte mínimo relativo vs. número de reglas encontradas

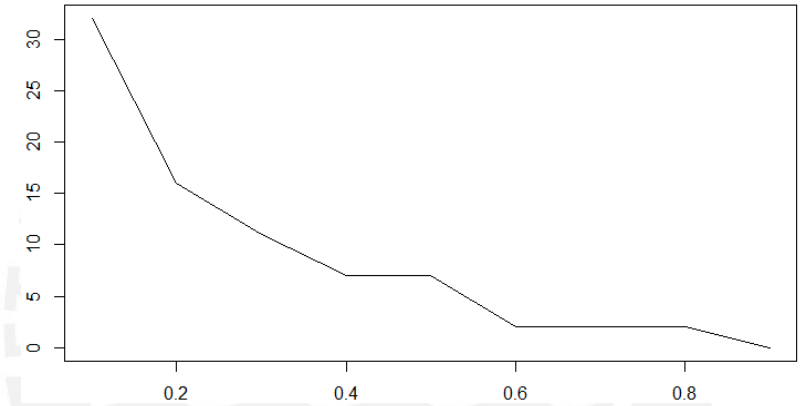


Figura 20: Soporte mínimo relativo vs. número de ciclos de procesamiento

Ítems precedentes	Soporte Absoluto del Precedente	Ítem consecuente	Soporte Absoluto del consecuente	Confianza
1=<modcourse<68, 1=<modresource<38	268	1=<actview<200	267	1
1=<modcourse<68	404	1=<actview<200	398	0.99
1=<modresource<38	289	1=<actview<200	281	0.97
1=<actview<200, 1=<modresource<38	281	1=<modcourse<68	267	0.95
1=<actview<200	421	1=<modcourse<68	398	0.95
1=<modresource<38	289	1=<modcourse<68	268	0.93
1=<modresource<38	289	1=<actview<200, 1=<modcourse<68	267	0.92

Tabla 14: Salida de reglas de asociación en Weka, minsup=0.4

En soportes menores otros elementos empiezan a aparecer junto con estos. El módulo carpeta (1=<modfolder<32) se encuentra cuando "1=<modcourse<68" está presente. Con soportes de 0.1 y 0.2 se encuentran ya otros elementos como "1=<modforum<46" y "1=<modurl<9" junto con los primeros tres elementos.

Las reglas de asociación y conjuntos frecuentes obtenidos en ambos casos no superaron la longitud de 5 elementos por conjunto, ni siquiera para los soportes mínimos más pequeños utilizados.

4.3.3 Patrones Secuenciales

Para la extracción de patrones secuenciales, se hizo uso de la librería SPMF, usando específicamente su implementación de PrefixSpan (Pei *et al*, 2004). Para su utilización, es necesario primero discretizar los datos de los registros utilizando rangos iguales. Al igual que para las reglas de asociación, se decidió partir los rangos de cada campo en 6. El número seis se eligió, nuevamente, considerando que los rangos de valores máximos para cada campo varían de 6 a 200. A cada conjunto de campo y rango se les dio una etiqueta única, eliminando los rangos que no contenían ningún elemento. La Tabla 15 muestra los primeros 10 elementos discretizados y sus significados.

Posteriormente, cada grupo de elementos (itemset) para un mismo usuario se agrupan secuencialmente. Esta agrupación supuso 492 grupos de secuencias de usuarios por semanas. Cabe recordar que para cada usuario sólo se utilizaba datos de un sólo curso.

Campo	Valores	Etiqueta
Actview	<40 & >=1	1
Actview	<80 & >=40	2
Actview	<120 & >=80	3
Actview	<160 & >=120	4
Actview	<200 & >=160	5
Actadd	<8 & >=1	6
Actadd	<16 & >=8	7
Actadd	<24 & >=16	8
actadd	<32 & >=24	9
actadd	<40 & >=32	10

Tabla 15: Primeros 10 elementos etiquetados para patrones secuenciales

La librería utilizada requiere además de los datos de entrada, indicarle un soporte mínimo, el cuál como se indicó anteriormente, indica el número mínimo de usuarios que contienen el patrón (frecuencia de aparición). Este soporte debe ingresarse de forma absoluta, por lo que debía ir desde el valor 1 hasta 492. Para analizar el comportamiento de los usuarios, se realizaron diferentes experimentaciones variando dicho soporte mínimo y los resultados cuantitativos de cada una se muestran en la Tabla 16.

Soporte Mínimo absoluto	Tiempo (ms)	Secuencias	Memoria (Mb)
490	64	3	32.53
475	293	12	62.52
450	297	12	62.51
425	1194	39	457.88
400	9200	124	690.12
375	57136	391	669.20
350	63411	516	772.94
325	342003	1884	1186.64
300	1633420	6747	1940.51
275	8214273	19112	1990.41
250	16576364	52874	1988.76
225	76597473	214653	1997.94
200	309486778	908555	1998.15

Tabla 16: Resultados cuantitativos de experimentación de patrones secuenciales

Las Figuras 21, 22 y 23 muestran la variación del tiempo de ejecución, número de secuencias encontradas y memoria utilizada dependiendo del soporte mínimo elegido. Se muestra que la mayor variación del tiempo y número de patrones se encuentra para un valor de soporte mínimo menor a 300, mientras que el consumo de memoria mantiene una alta variación entre el rango de 300-450 de soporte mínimo para luego estabilizarse.

Esta variación de memoria resulta en una limitación de las experimentaciones. No se realizaron experimentaciones con soportes menores a 200 porque el costo computacional se volvió muy alto. La prueba de 200 demoró más de tres días en completarse.

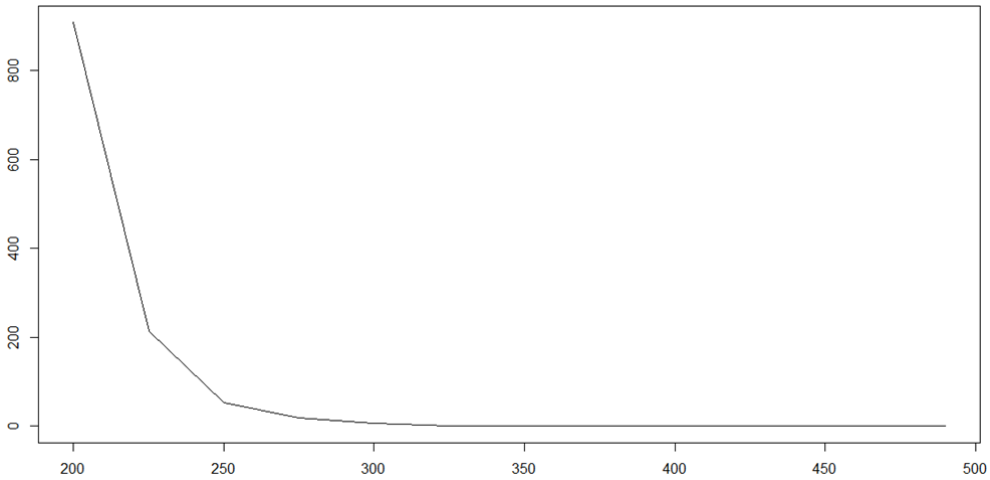


Figura 21: Soporte mínimo absoluto vs. número de patrones encontrados (en miles)

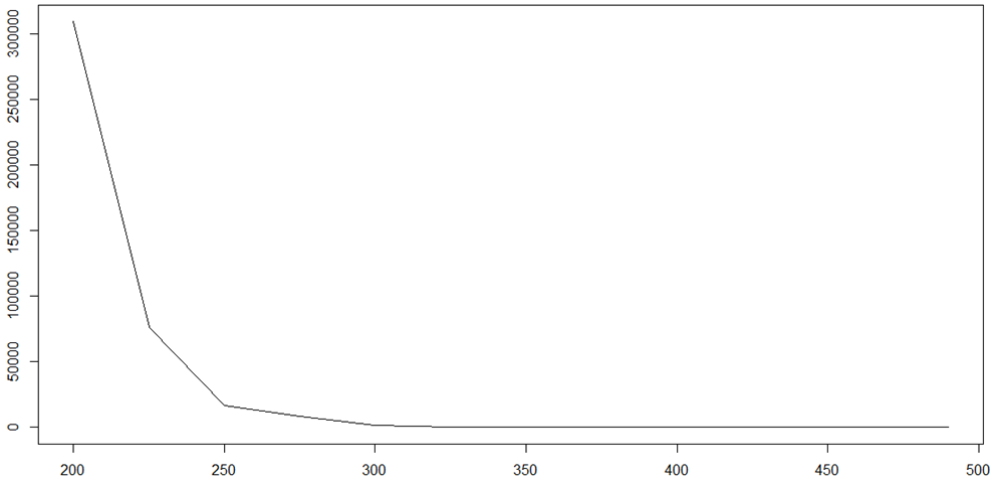


Figura 22: Soporte mínimo absoluto vs. tiempo en segundos

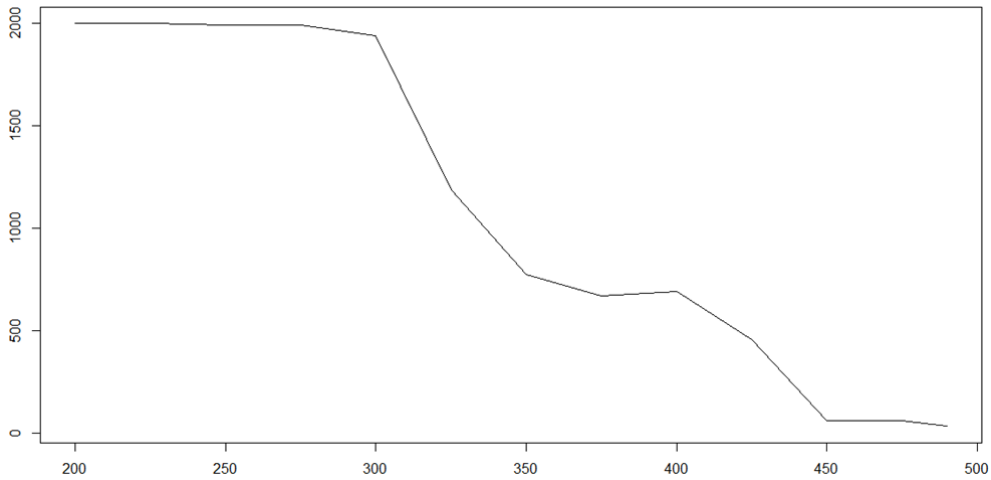


Figura 23: Soporte mínimo absoluto vs. memoria (en Mb)

Los patrones que más se encontraron usando diferentes soportes mínimos, eran los relacionados a la actividad "ver", así como en los módulos de "recurso" (*resource*) y "curso" (*course*) los cuales se mantuvieron sin embargo en los rangos de menos de 50 *clics* por semana. La repetición del patrón "modcourse <50" y "actview <40" se repite de semana en semana en diferentes combinaciones exclusivamente hasta bajar al soporte mínimo absoluto a 425. La Tabla 17 incluye todos los patrones encontrados en esta experimentación (soporte absoluto = 425). El elemento **"1"** se refiere a **"1=<actview <40"** (la acción de ver un elemento en la plataforma) y el elemento **"40"** a **"1=<modcourse<50"** (el interactuar con el curso en general, no con sus elementos).

Se muestra que la secuencia más larga no supera los tres elementos (semanas) y que en algunos casos sólo se expresa uno de los elementos como común. Además, se nota la regularidad de que si un usuario ingresa a la plataforma sólo a visualizar algún elemento, lo más probable es que sólo sea a mirar el contenido en general del curso.

Con menores soportes, aparecen otros patrones en repetición. Por ejemplo, el elemento "modresource <30" (entre uno y 30 clics en un recurso) empieza a formar parte de las secuencias junto con los otros dos elementos ya indicados al bajar al soporte mínimo 375. Así mismo, la longitud de las secuencias empiezan a crecer dando patrones de 5 elementos (con soporte 375), 6 elementos (con soporte 325) y 7 elementos (con soporte 300). Ningún otro elemento más que los tres mencionados (modresource <30, actview <40 y modcourse <50) fue encontrado utilizando los soporte mínimos de las experimentaciones.

La falta de aparición de otros elementos frecuentes muestra que aunque existe variación en los atributos, estás variaciones no se repiten el tiempo con la suficiente frecuencia para ser extraídos. Sin embargo, como se ha podido bajar el soporte mínimo poco menos del 50%, puede que existan pero supone un costo demasiado alto.

Se realizó otras experimentaciones utilizando CloSpan (Yan, Han & Afshar, 2003), a fin de revisar si con patrones secuenciales cerrados, es decir, que no se incluyan entre sí. El resumen de las experimentaciones puede encontrarse en la Tabla 18. El mínimo soporte relativo que se consiguió experimentar fue 35, dado que el consumo de memoria evitaba el completar la experimentación. Se realizó tres experimentaciones, buscando encontrar soportes distintos a los patrones secuenciales ya encontrados.

Los resultados muestran que el tiempo y memoria crecen exponencialmente, por lo que no se realizaron más experimentaciones. Así mismo, revisando los resultados generados, el único patrón distinto a los tres antes mencionados (0 < modresource <

30, 0 < actview < 40 y 0 < modcourse <50) fue 0 < modfolder <15, el cual se refiere a las carpetas que existen dentro de los cursos. No se ganó mayor información utilizando este algoritmo. La Figura 24 muestra la variación en memoria de las experimentaciones.

Conjuntos	Longitud de secuencia	Soporte		Conjuntos	Longitud de secuencia	Soporte
1 //	1	492		1 40 // 1 40 // 1 40 //	3	441
1 // 1 //	2	476		1 40 // 1 40 // 40 //	3	442
1 // 1 // 1 //	3	441		1 40 // 40 //	2	475
1 // 1 // 1 40 //	3	441		1 40 // 40 // 1 //	3	441
1 // 1 // 40 //	3	442		1 40 // 40 // 1 40 //	3	441
1 // 1 40 //	2	475		1 40 // 40 // 40 //	3	442
1 // 1 40 // 1 //	3	441		40 //	1	492
1 // 1 40 // 1 40 //	3	441		40 // 1 //	2	476
1 // 1 40 // 40 //	3	442		40 // 1 // 1 //	3	441
1 // 40 //	2	475		40 // 1 // 1 40 //	3	441
1 // 40 // 1 //	3	441		40 // 1 // 40 //	3	442
1 // 40 // 1 40 //	3	441		40 // 1 40//	2	475
1 // 40 // 40 //	3	442		40 // 1 40 // 1 //	3	441
1 40 //	1	492		40 // 1 40 // 1 40 //	3	441
1 40 // 1 //	2	476		40 // 1 40 // 40 //	3	442
1 40 // 1// 1 //	3	441		40 // 40 //	2	475
1 40 // 1// 1 40 //	3	441		40 // 40 // 1 //	3	441
1 40 // 1// 40 //	3	442		40 // 40 // 1 40 //	3	441
1 40 // 1 40 //	2	475		40 // 40 // 40 //	3	442
1 40 // 1 40 // 1 //	3	441				

Tabla 17: Patrones secuenciales obtenidos por SPMF (minsup=425)

Soporte Mínimo Relativo	Tiempo (ms)	Secuencias	Memoria (Mb)
0.5	17173	2453	392.24
0.4	135219	12003	706.11
0.35	441116	33419	1565.93

Tabla 18: Resultados cuantitativos de experimentación de patrones secuenciales cerrados

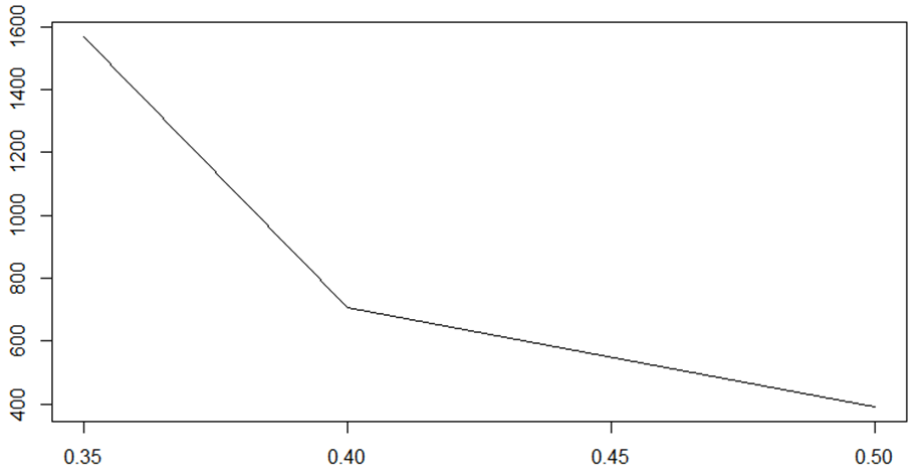


Figura 24: Soporte mínimo absoluto vs. memoria (en Mb) con CloSpan

Capítulo 5. Análisis de Resultados: Interpretación y validación

En esta sección se discutirá los resultados obtenidos en las experimentaciones detalladas en la sección anterior.

5.1 Clusterización

Las experimentaciones de clusterización resultaron en tres clústeres bastante sesgados cuando se utilizó el primer algoritmo (EM), las cuales se muestran en las Figuras 10-13. Más bien, se concluye que el segundo algoritmo, al buscar balancear los clústeres, genera grupos artificiales que no necesariamente caracterizan a los usuarios.

Tomando el primer resultado, entonces, interpretamos que uno de los grupos debe consistir en usuarios docentes o asistentes de docencia, dado que la acción de "agregar" elementos es uno de los atributos que más lo distingue del resto. El clúster de mayor población a su vez, puede ser considerado de alumnos de comportamiento regular, teniendo en cuenta además, que mantienen valores bajos de acciones en general.

El tercer y último clúster, sin embargo, genera cierta incógnita. Primero, porque cuenta sólo con dos elementos dentro de la población y por otro porque no es especialmente distinguible de los demás. Buscando entre elementos que los caracterizan, se encuentra su presencia como usuarios que realizaron la acción "editar", la cual es exclusiva del clúster de docentes. Graficamos dicha relación en la Figura 25. Por tanto, se interpreta que este pequeño clúster caracteriza a docentes (o asistentes de docencia) que no tienen tanta variabilidad o cantidad de acciones como ellos pero que cuentan con un rol similar a ellos.

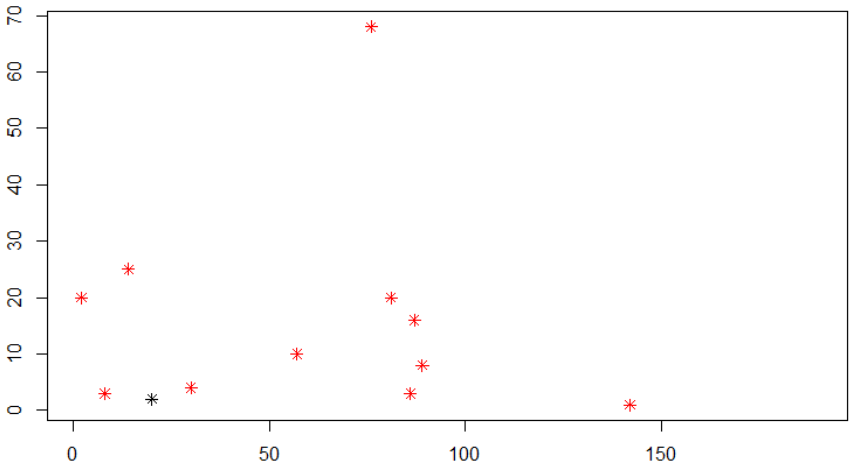


Figura 25: Actividad "editar" Vs. módulo "recurso", clúster por Algoritmo EM

5.2 Reglas de Asociación e Itemsets Frecuentes

Las experimentaciones de reglas de asociación e itemsets frecuentes mostraron los patrones más comunes dentro de los datos y a su vez la asociación entre sus apariciones. No es una sorpresa encontrar que los elementos más frecuentes son la acción de “ver” un elemento en la plataforma y que los elementos sean el curso en general y los recursos.

Para visualizar la relación entre dichos elementos, se ha generado un grafo (transformando a JSON la salida de Weka) que muestra las relaciones encontradas usando un soporte relativo de 0.1, utilizando un algoritmo de fuerza. La semántica visual que se ha utilizado es la siguiente:

- Los colores de nodos especifican si se trata de un nodo que agrupa 1 (azul), 2 (celeste) o 3 (naranja) atributos.
- La distancia entre nodos es inversamente proporcional a la confianza de la relación. Mientras más confianza, más corta la distancia.
- El tamaño de los nodos especifica la frecuencia de aparición de dicho atributo o grupo de atributos.
- La carga de los nodos es relativa a si son grupos de 1, 2 o 3 atributos. Se intenta que no se solapen entre ellos.
- Cada nodo incluye una etiqueta especificando el atributo: son rangos de aparición de cada elemento.

El gráfico final puede visualizarse en la Figura 26. En éste, se nota que los centros de las relaciones dependen casi exclusivamente de los rangos mínimos de la acción ver (actview) y de los módulos de curso y recurso. A su alrededor se encuentran otros tipos de elementos, como el módulo de carpeta, url y foro. Cabe resaltar que ninguna otra acción aparece en las reglas.

Esta información coincide con los análisis iniciales, en los que se demostró cuáles eran los módulos y acciones más comunes. Sin embargo, el gráfico permite notar cuáles de estos módulos están más relacionados tanto con la acción ver como con otros módulos. Algunos ejemplos son el nodo que une los módulos recurso y curso (modresource y modcourse) el cual es consecuente de las acciones de ver, el cual tiene mucha más frecuencia comparándolo con el nodo que relaciona el módulo de recurso y de url. Otro punto importante está asociado a las relaciones, tomando como ejemplo la relación de la acción ver cuya distancia es más corta (y por tanto más cercana) al módulo de recurso que al módulo de folder.

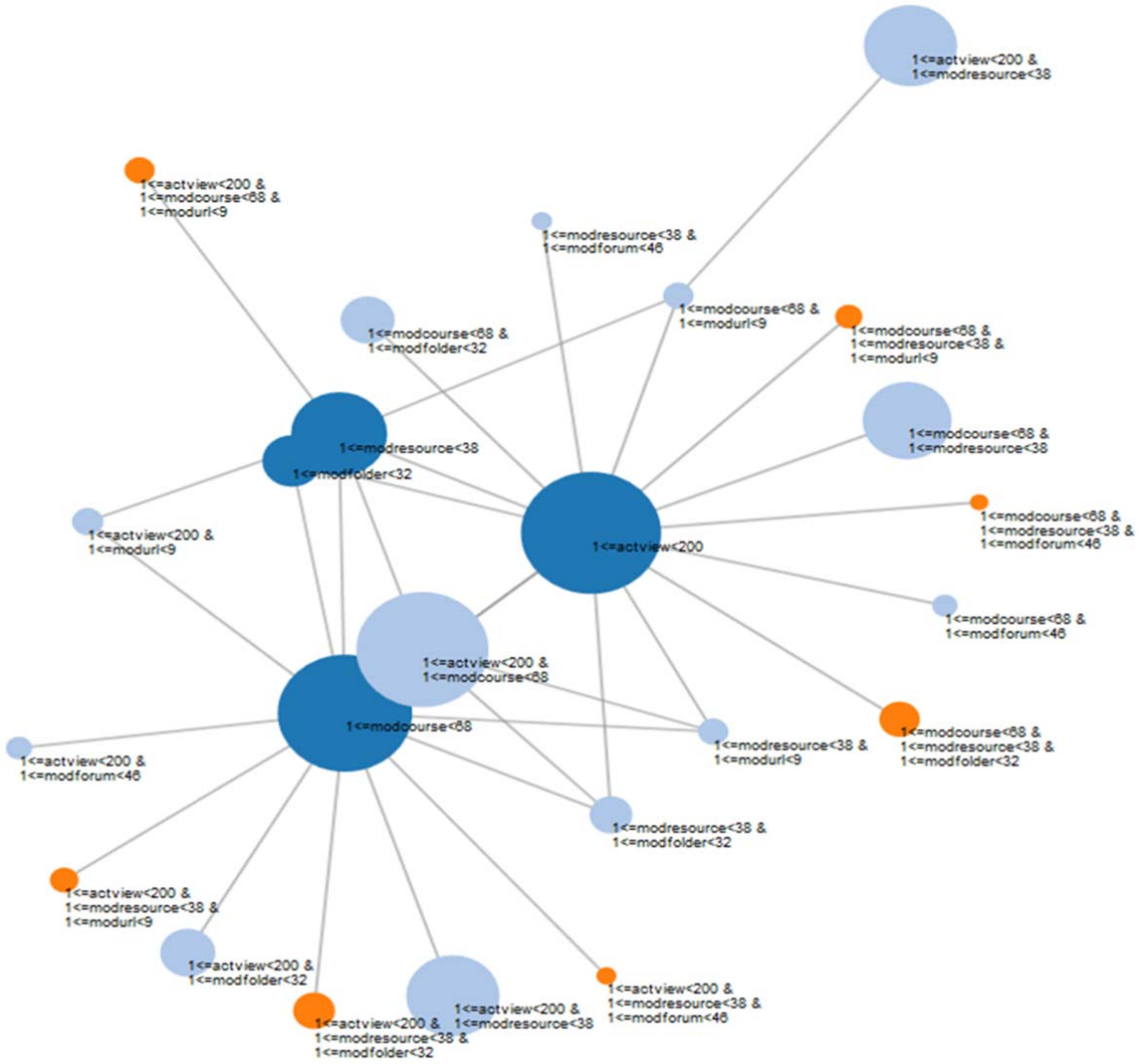


Figura 26: Grafo de fuerza, representando las reglas de asociación halladas con Apriori usando Weka, con soporte mínimo relativo de 0.1

Muchas otras relaciones, como por ejemplo la frecuencia de la aparición conjunta de recurso, curso y foro, pueden ser inferidas y visualizadas en el gráfico propuesto, lo cual lo hace una herramienta útil para la retroalimentación de los usuarios de la plataforma.

5.3 Patrones Secuenciales

Las experimentaciones usando patrones secuenciales mostraron poca variabilidad, siendo sólo tres elementos los que se encontraron recurrentemente: modcourse, modresource y actview. Se decidió crear un pequeño grafo con los resultados más cortos (soporte mínimo absoluto: 425) utilizando la siguiente semántica.

- Los colores de nodos especifican si se trata de un nodo que aparece primero en el patrón (azul), segundo (celeste) o tercero y último (naranja).
- La distancia entre nodos es inversamente proporcional a la frecuencia del patrón. Mientras más frecuencia, más corta la distancia. Se incluye dirección.
- El tamaño de los nodos y su carga es uniforme.
- Cada nodo incluye una etiqueta especificando el atributo: rangos de aparición de cada elemento.

El gráfico se visualiza en la Figura 27. En ella se notan las combinatorias de apariciones de los tres elementos, siendo especialmente notorias las distancias relativas a la aparición del patrón. Por ejemplo, dada una aparición de 1<=modcourse<50 en un primer espacio temporal, es más probable que se repita dicho patrón que la aparición del elemento 1<=actview<40. Sin embargo, al ser el soporte relativo bastante alto, las distancias no varían demasiado entre sí.

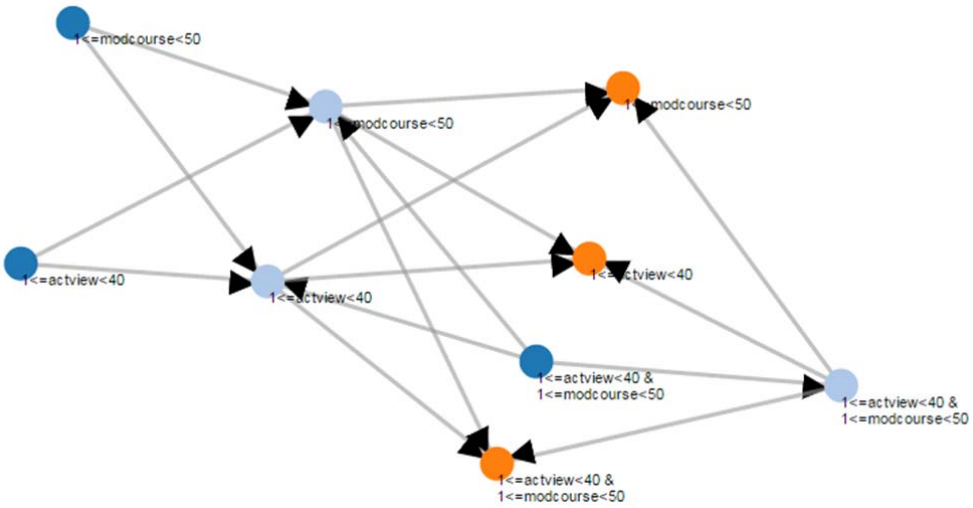


Figura 27: Grafo de fuerza, representando patrones secuenciales con soporte mínimo 425/492.

Adicionalmente al grafo, es necesario notar que en las múltiples experimentaciones, los patrones que se formaron mantenían cierta regularidad al repetir continuamente los rangos de acciones realizadas y/o módulos accedidos. Eso se integra a la necesidad de bajar progresivamente los límites del soporte mínimo para encontrar otro tipo de atributos, por lo que se concluye que el comportamiento de los usuarios es **regular en su mayor parte**. Esta última observación marca una característica a revisar para los que utilizan la plataforma, ya que encontrar este tipo de patrón, permite caracterizar a los usuarios y preparar los contenidos considerando este tipo de usuarios.

Capítulo 6. Conclusiones, limitaciones y trabajos futuros

6.1 Conclusiones

El estudio surgió como una necesidad de analizar los datos acumulados de la plataforma Paideia utilizada en la Pontificia Universidad Católica del Perú. Para ello, se usó la metodología KDD, utilizando distintos tipos de algoritmos de minería de datos a fin de extraer patrones.

Como conclusión general, podemos afirmar que es posible extraer patrones semánticamente distintos de los datos de los logs de una plataforma educativa, dados los resultados explicados en los capítulos anteriores,

En específico:

- Se generó un conjunto de datos basado en el contenido de la tabla de log, de forma que pudo aplicarse minería de datos a este.
- La clusterización permitió agrupar usuarios de acuerdo a los tipos de acciones y módulos que han accedido. Los resultados permitieron identificar tres grupos, uno de los cuales tenía los permisos de profesor.
- Las reglas de asociación permitieron descubrir correlaciones entre elementos. Estos resultados permitieron identificar cuándo aparecen juntos los diferentes atributos, los cuales se centraron en las acciones de visualizar y el módulo de curso.
- Los patrones secuenciales permitieron evaluar cómo se comportan los usuarios a través del tiempo. Una conclusión que se pudo extraer a partir de estos patrones, es que el comportamiento de los usuarios de Paideia no varía en el tiempo. Aquellos que realizan ciertas acciones con ciertos módulos, comúnmente siguen haciéndolo en las diferentes unidades de tiempo frecuentemente.
- La visualización utilizada fue de grafos no dirigidos para mostrar las regalas de asociación y un pequeño grafo dirigido para expresar parte de los resultados de patrones secuenciales.

6.2 Limitaciones de la investigación

La presente investigación se trabajó a modo de un estudio exploratorio, por lo que utilizar todos los datos disponibles no era el objetivo del mismo. Sin embargo, la selección de los datos puede haber sesgado los patrones finalmente encontrados, ya que se eligió de forma arbitraria un protocolo de selección de la muestra.

Los resultados más específicos obtenidos de la investigación no deben tomarse como absolutos, sino más bien como ejemplos de resultados posibles al usar los registros de la plataforma para extraer patrones. Para un análisis más minucioso de la

validez de los resultados, se requeriría seleccionar las características que las muestras deben cumplir para ser estadísticamente significativas.

Otra limitación fue el poder computacional con el que se contaba, ya que cuando se intentó a llegar a bajos soportes mínimo, algunos de los algoritmos tomaban mucho tiempo o consumían demasiada memoria. Sin embargo, los soporte más altos son aquellos que permiten encontrar los atributos más comunes.

6.3 Trabajos Futuros

Los patrones semánticamente distintos han servido para interpretar los resultados esperados en este proyecto. Como trabajos futuros se podría ampliar el número de datos utilizados, o seleccionarlos por curso, y hacer el mismo análisis de extracción de patrones. Es, sin embargo, importante considerar que para realizar estos nuevos estudios aumentará el costo computacional, por lo que se debe equilibrar el costo-beneficio.

Por otro lado, sería posible expandir el análisis utilizando la información de IP, lo cual nos puede brindar la ciudad en la que los usuarios se conectaron. Agregar dichos datos a los anteriores podría ayudar a encontrar patrones diferentes según la ubicación.

Otra extensión de este trabajo sería el utilizar los grafos para visualizar todas las salidas de patrones secuenciales, incluso las más extensas. En la presente investigación se tomó de ejemplo una de las salidas más cortas, pero otros trabajos podrían incluir un esfuerzo especial en la visualización y automatizar de cierta forma la transformación de los datos a este tipo de estructura.

Finalmente, se han realizado otro tipo de estudios (Cohn-Muroy, Flores-Lafosse & Villanueva, 2015) en la plataforma en el año 2015, pero más enfocados en la percepción de los alumnos. El unir este estudio con ese tipo de investigaciones podría redondear más la retroalimentación a los docentes, dando una visión más cualitativa del uso de la plataforma.

Referencias bibliográficas

Abdullah, M., Alqahtani, A., Aljabri, J., Altowirgi, R., & Fallatah, R. (2015). Learning Style Classification Based on Student's Behavior in Moodle Learning Management System. http://doi.org/10.14738/tmlai.31.868

Abdullah, Z., Herawan, T., Chiroma, H., & Deris, M. M. (2014). A Sequential Data Preprocessing Tool for Data Mining. In *Computational Science and Its Applications--ICCSA 2014* (pp. 734–746). Springer.

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*. http://doi.org/10.1109/ICDE.1995.380415

Agrawal, R., & Srikant, R. (1994, September) *Fast algorithms for mining association rules*. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Blot, G., Saurel, P., & Rousseaux, F. (2014). Pattern Discovery in E-learning Courses : a Time- based Approach.

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 - LAK '14* (pp. 11–15). New York, New York, USA: ACM Press. http://doi.org/10.1145/2567574.2567604

Castro F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying Data Mining Techniques to e-Learning Problems\r. *Studies in Computational Intelligence (SCI)*, *62*(2007), 183–221.

Cohn-Muroy, D., Flores-Lafosse, N., & Villanueva, V. (2015). *Percepción del uso de una plataforma virtual de aprendizaje colaborativo en una universidad peruana: Estudio de Caso*. Anais temporários do LACLO 2015, 10(1), 81.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the royal statistical society. Series B (methodological), 1-38

Dimopoulos, I., Petropoulou, O., & Retalis, S. (2013). Assessing students' performance using the learning analytics enriched rubrics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13* (p. 195). New York, New York, USA: ACM Press. http://doi.org/10.1145/2460296.2460335

Falakmasir, M. H., Moaven, S., Abolhassani, H., & Habibi, J. (2010). Business intelligence in e-learning: (case study on the Iran university of science and technology dataset). *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Filva, D. A., Guerrero, M. J. C., & Forment, M. A. (2014). Google analytics for time behavior measurement in Moodle. In *2014 9th Iberian Conference on*

Information Systems and Technologies (CISTI) (pp. 1–6). IEEE. http://doi.org/10.1109/CISTI.2014.6877095

Florian, B., Glahn, C., Drachsler, H., Specht, M., & Fabregat Gesa, R. (2011). *Activity-based learner-models for learner monitoring and recommendations in Moodle*. (C. D. Kloos, D. Gillet, R. M. Crespo García, F. Wild, & M. Wolpers, Eds.)*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6964). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-23985-4

Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu., C. & Tseng, V. S. (2014). *SPMF: a Java Open-Source Pattern Mining Library*. Journal of Machine Learning Research (JMLR), 15: 3389-3393.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I-H. (2009). *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

Han, J. (2014). *Data Mining : Concepts and Techniques*.

Kapros, E., & Peirce, N. (2014). *Learning and Collaboration Technologies. Designing and Developing Novel Learning Experiences*. (P. Zaphiris & A. Ioannou, Eds.)*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8523). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-07482-5

Kato, T., & Ishikawa, T. (2013). *Detection and Presentation of Failure of Learning from Quiz Responses in Course Management Systems*. (T. Yoshida, G. Kou, A. Skowron, J. Cao, H. Hacid, & N. Zhong, Eds.)*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8210). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-02750-0

Kearns, M., Mansour, Y., & Ng, A. Y. (1998). *An information-theoretic analysis of hard and soft assignment methods for clustering*. In *Learning in graphical models* (pp. 495-520). Springer Netherlands.

Kotsiantis, S., Tselios, N., Filippidi, A., & Komis, V. (2013). Using learning analytics to identify successful learners in a blended learning course. *International Journal of Technology Enhanced Learning*, *5*(2), 133. http://doi.org/10.1504/IJTEL.2013.059088

Krpan, D., & Stankov, S. (2012). Educational Data Mining for Grouping Students in E-learning System. *Proceedings of the Iti 2012 34th International Conference on Information Technology Interfaces*, 207–212. http://doi.org/10.2498/iti.2012.0470

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Mahmoud, M. (2008). Data Mining : Concepts and Techniques Summer Course 1429 H By.

Martinovi, J. (2012). Dynamic Time Warping in Analysis of Student Dynamic Time Warping in Analysis of Student Behavioral Patterns Behavioral Patterns, 49–59.

Masseglia, F., Teisseire, M., & Poncelet, P. (2005). Sequential Pattern Mining, 1–5. http://doi.org/10.1002/pmic.200700657

Miccoli, F (2013) Data visualization, machine learning. https://www.mysciencework.com/news/9823/data-visualization-machine-learning

Monk, D. (2005). Using data mining for e-Learning decision making. *Electronic Journal of E-Learning*, *3*(1), 41–54.

Morales, E., & Gonzales, J. (2013). Reglas de asociación.

Nagi, K., & Suesawaluk, P. (2008). Research analysis of moodle reports to gauge the level of interactivity in elearning courses at Assumption University, Thailand. *Proceedings of the International Conference on Computer and Communication Engineering 2008, ICCCE08: Global Links for Human Development*, 772–776. http://doi.org/10.1109/ICCCE.2008.4580710

Nebić, Z., & Mahnič, V. (2010). Data warehouse for an e-learning platform, (Volume II), 415–420.

Oficina de Validación y Análisis de Tecnologías para la Educación - VATE (2013). Infografía innovación educativa con TIC http://vate.pucp.edu.pe/proyectos/otros/infografia-innovacion-educativa-con-tic/ Visitada: 14/11/2015

Ogashiwa, K., Hamamoto, Y., Wang, Y., Kariya, J., & Ogawara, K. (2013). Graphical tool for formative assessment with the Moodle quiz module. In *Work-in-Progress Poster (WIPP) Proceedings of the 21st International Conference on Computers in Education, ICCE 2013* (pp. 15–17). UHAMKA PRESS.

Palace, B (1996) Data Mining http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., & Hsu, M. C. (2004). *Mining sequential patterns by pattern-growth: The prefixspan approach.* Knowledge and Data Engineering, IEEE Transactions on, 16(11), 1424-1440.

Pelleg, D., & Moore, A. W. (2000, June). *X-means: Extending K-means with Efficient Estimation of the Number of Clusters*. In ICML (pp. 727-734).

Petropoulou, O., Kasimatis, K., Dimopoulos, I., & Retalis, S. (2014). LAe-R: A new learning analytics tool in moodle for assessing students' performance. *Bulletin of the Technical Committee on Learning Technology*, *16*(1), 2–5. http://doi.org/10.1010101

Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135–146. http://doi.org/10.1002/cae.20456

Romero, C., González, P., Ventura, S., del Jesus, M. J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. *Expert Systems with Applications*, *36*(2 PART 1), 1632–1644. http://doi.org/10.1016/j.eswa.2007.11.026

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

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

Serrano-Laguna, Á., Torrente, J., Maneroa, B., Del Blanco, Á., Borro-Escribano, B., Martínez-Ortiz, I., Fernández-Manjón, B. (2014). *Learning Analytics and Educational Games: Lessons Learned from Practical Experience*. (A. De Gloria, Ed.) *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8605). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-12157-4

University of Hamburg (2015). Statistical analysis of machine learning algorithms

https://www.informatik.uni-hamburg.de/ML/contents/research/index.shtml

Visitada: 14/11/2015

Vishwakarma, A. (2014). A Survey on Web Log Mining Pattern Discovery, *5*(6), 7022–7031.

Xiong, L. (2008). Data Mining : Concepts and Techniques Cluster Analysis.

Yan, X., Han, J., & Afshar, R. (2003, May). *CloSpan: Mining closed sequential patterns in large datasets*. In In SDM (pp. 166-177).

Zhu, H., Zhang, X., Wang, X., Chen, Y., & Zeng, B. (2014). A Case Study of Learning Action and Emotion from a Perspective of Learning Analytics. In *2014 IEEE 17th International Conference on Computational Science and Engineering* (pp. 420–424). IEEE. http://doi.org/10.1109/CSE.2014.105