Minería de uso Web: Creación de perfiles, para la recomendación y optimización del uso en plataformas virtuales educativas.

Corso, Cynthia Lorena
Universidad Tecnológica Nacional – Facultad Regional Córdoba
cynthial@bbs.frc.utn.edu.ar
Alfaro, Sofía Lorena
soft@bbs.frc.utn.edu.ar
Universidad Tecnológica Nacional – Facultad Regional Córdoba

Abstract

Este trabajo de investigación tiene como objetivo la presentación de un caso de aplicación relacionado con la minería web, más específicamente a la minería de uso web, aplicado al ámbito de la educación.
En él se expone una metodología para la aplicación de la minería de uso web, centrándonos en la creación de perfiles, como medio de retroalimentación de los sistemas de recomendación implementados en el ámbito de la enseñanza basada en el uso de plataformas virtuales.

Palabras Clave
Minería Web, Sistemas de recomendación, Creación de perfiles, Moodle.

1. Introducción

En los últimos años, el crecimiento tecnológico y la expansión del uso de internet en todos los ámbitos, ha generado la necesidad de almacenar volúmenes significativos de información. Aunque pueda llegar a suponerse, que el disponer de mayor cantidad de datos almacenados, significa un aporte de conocimiento no siempre es así. La dificultad se plantea, que con las técnicas clásicas de recuperación de información, no cubren las expectativas para la gran cantidad de datos almacenados.

En los últimos años, han aparecido nuevas técnicas de procesamiento con mayor potencialidad y que dan respuesta a esta problemática.
La minería de datos puede definirse como la extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de los datos.

El objetivo principal de este trabajo es profundizar en una de las extensiones de la minería de datos, aplicadas al ámbito del uso de internet.

En la actualidad la gran mayoría de las empresas, invierten en la realización de sitios web, ya sea para su propia gestión como así también la interacción con sus clientes. En la gran mayoría de los casos, no se considera un factor fundamental que es el proceso posterior a la implementación, como el mantenimiento y los posibles planes de mejora.
El hecho de contemplar estos aspectos, facilita que por ejemplo los potenciales clientes puedan encontrar de manera rápida y simple lo que buscan o servicios complementarios.

En el contexto educativo la aparición de las plataformas virtuales educativas ha reunido en una sola alternativa, funcionalidades que facilitan la administración y organización de contenidos a medida de los alumnos.
En la Universidad Tecnológica Nacional F.R.C aproximadamente hace unos 5 años, se ha incorporado a las asignaturas de la carrera de Ingeniería en Sistemas de Información. La gran mayoría de las cátedras utiliza funcionalidades de Moodle, con los siguientes objetivos:

- Publicación de Contenidos Curriculares.
- Publicación de Novedades de la Asignatura.
- Foros.
- Automatización de Trabajos Prácticos.
- Automatización de Examen Parciales Teóricos/Prácticos.
- Automatización de actividades de autoevaluación.

Esta plataforma educativa proporciona tanto a docentes como alumnos una importante variedad de recursos, lo que en muchas ocasiones esto suele ser una desventaja. En algunas ocasiones el alumno puede estar desorientando a la hora de seleccionar los recursos diseñados, que la gran mayoría son numerosos. Esta dificultad requiere una revisión de cómo el docente ha reestructurado y diseñado la asignatura en la plataforma. Es importante conocer si el docente conoce cuales son las características del grupo, con la finalidad que el diseño resultante se adapte fácilmente a los alumnos. [1]

Los sistemas de recomendación son estrategias de personalización utilizadas para sugerir productos y/o contenido a los visitantes de sitios Web a partir de sus preferencias. El comportamiento de navegación que presentan los usuarios es un indicador de estas preferencias y es la base para extraer patrones de uso frecuente, también llamados perfiles de usuario. [*]

Atendiendo a los aspectos que se desean investigar, la minería web se clasifica en tres ramas:

Minería de contenido web: Se centra en los datos reales que contienen las páginas, es decir el contenido, estos datos consisten generalmente en textos y gráficos. Y se puede obtener datos que acerca de la forma de escribir que elementos o recursos son más atractivos para el usuario, de si la catalogación que usamos sirve para mejorar un ranking, si los temas que se tratan interesan o no.

Minería de estructura web: En esta variante de web mining, la información a emplear es aquella que describe la organización del contenido de la página web. La información de la estructura interna de una página incluye por ejemplo el arreglo de las varias etiquetas HTML o de XML dentro de esta. Esto se puede representar como una estructura de árbol en donde una etiqueta (HTML) se convierte en la raíz de este. La clase principal de información en la estructura interna de una página son los "hyper-links" que conectan una página con otra. El tipo de información acerca de si los usuarios encuentran la información, si la estructura de sitio es demasiado ancha o demasiado profunda, si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación se entiende, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central.

Minería de uso web: La minería de uso en la web consiste en la aplicación de técnicas de minería de datos para descubrir los patrones de uso de la información web con el objetivo de entender y satisfacer las necesidades de los usuarios [2].
Las principales técnicas usadas son de análisis estadístico, reglas de asociación, agrupación de ítems, clasificación y modelos de dependencia aplicadas a las bitácoras de los servidores web y secuencia de páginas visitadas para realizar transacciones. [3]
Los aspectos en los que se enfoca están orientados a descubrir el uso de las páginas web, direcciones IP, referencias a páginas, la fecha y la hora de accesos, datos que proporcionan la información demográfica sobre los usuarios del Sitio Web. La fuente de datos típica en la minería de uso web son: logs de acceso en servidores, proxies, agentes y cookies.

Nuestro trabajo nos enfocaremos en la minería de uso web, aplicado a un caso práctico donde se expone una propuesta metodológica para obtener patrones de uso.

2. Elementos del Trabajo y Metodología

Atendiendo a los objetivos que este trabajo pretende alcanzar, el mismo se focaliza en la rama de la minería web más precisamente sobre la minería de uso web. Las etapas planificadas para lograr los objetivos establecidos en el apartado anterior son:

Preprocesado: Esta etapa incluirá un conjunto de tareas como limpieza de datos, identificación de las páginas visitadas, de usuarios y sesiones.
La identificación de los recursos visitados nos permitirá obtener conocimiento relacionado con el tipo de actividad que los alumnos desarrollan en el uso de la plataforma.
Si bien existen diversas técnicas para identificar usuarios como creación de perfiles o cookies o la identificación de direcciones IP cada una de ellas con sus ventajas y desventajas, nosotros utilizaremos un recurso de Administración llamado "Informe", que está incorporado a la plataforma virtual Moodle, solamente es accedido por el perfil Docente. De esta opción existen diversas variantes, la que nos interesa para nuestro estudio es "Estadística".
Esta herramienta nos ofrece la posibilidad de seleccionar la asignatura, el tipo de Informe es decir si lo que se necesita disponer una vista de la actividad del Administrador, Profesor, Alumno o Invitados en nuestro caso nos interesa la vista de los alumnos. Además otra posibilidad que nos brinda es obtener información histórica de hasta dos años atrás.
A continuación se muestra un caso concreto de la actividad de los alumnos 9 meses atrás.

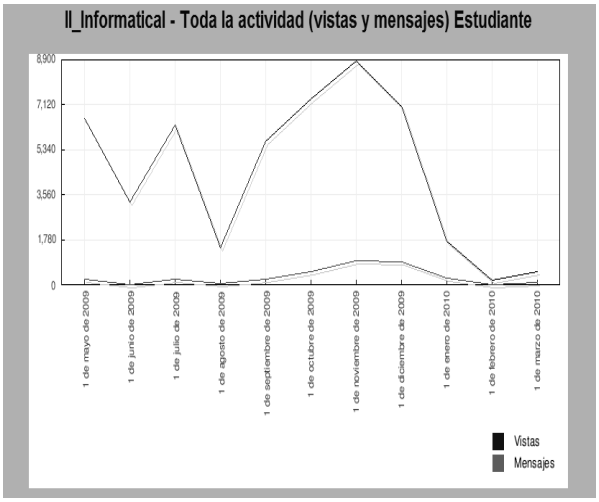

Figura1. Representación gráfica de la actividad de los alumnos.

Fecha	Dirección IP	Nombre completo	Acción	Información
lun 30 de noviembre de 2009, 22:11	190.230.170.3	Federico Herrero	resource view	Guía de ACCESS con Bases de Datos de ejemplos
lun 30 de noviembre de 2009, 22:11	190.230.170.3	Federico Herrero	course view	Informatica I
lun 30 de noviembre de 2009, 21:48	190.230.168.41	Lucas Alejandro Muñoz	course view	Informatica I
lun 30 de noviembre de 2009, 21:30	200.117.208.59	Agustín Andrés Pintado	quiz view all	
lun 30 de noviembre de 2009, 21:18	190.228.229.87	Federico Herrero	course view	Informatica I
lun 30 de noviembre de 2009, 21:12	190.227.249.220	Neri Gustavo Herrador Rojo	course view	Informatica I
lun 30 de noviembre de 2009, 20:57	201.252.163.105	Susana Virginia Benito	resource view	Información sobre SAP R/3
lun 30 de noviembre de 2009, 20:51	201.252.163.105	Susana Virginia Benito	resource view	Link - Administración de Proyectos de Sistemas
lun 30 de noviembre de 2009, 20:50	201.252.163.105	Susana Virginia Benito	resource view	Link - Administración de Proyectos de Sistemas
lun 30 de noviembre de 2009, 20:47	201.252.163.105	Susana Virginia Benito	resource view	Guía de ACCESS con Bases de Datos de ejemplos

Figura2. Información de la sesión de los alumnos con actividad en esa fecha.

Como se puede visualizar en la Figura 2, no solo se obtiene información de manera gráfica como lo que se muestra en la Figura 1. Además se presenta un listado con toda la actividad de los alumnos relacionada con el uso de la plataforma Moodle en el periodo considerado. Los datos que se visualizan son:

- Fecha de la sesión,
- Dirección IP,
- Nombre del alumno,
- Acción que está relacionada con el tipo de actividad que hizo el alumno
- Información específica relacionada con el recurso accedido (foro, encuesta, actividad evaluación etc.).

Limpieza y Transformación de Datos: Si bien todos los datos proporcionados por la herramienta son pertinentes para nuestro estudio, algunos de los datos no lo consideramos para nuestro estudio.
Por ejemplo la dirección IP no nos aportará ningún aspecto significativo en nuestro estudio, como así también el nombre del alumno.
Para facilitar el tratamiento de estos ítems considerados y prepararlo para que pueda ser aplicado el algoritmo KMeans, el campo Información específica que denota el tipo de actividad que accedió el alumno, se transformación a un tipo de variable cualitativa nominal.

Los valores posibles son:
G: representa acceso a información general relacionada con el curso.
F: representa acceso a los foros de la asignatura.
O: representa acceso a las evaluaciones formativas de la asignatura.
E: representa acceso al material y bibliografía de la asignatura.

Para poder llevar a cabo esta tarea, la alternativa de software seleccionada para el aprendizaje automático será Weka. Esta herramienta ha sido desarrollada bajo la concepción de software libre.
Es una herramienta muy potente, ya que incluye diferentes módulos desde la preparación de los datos, hasta la incorporación de diversos algoritmos relacionados con las técnicas de clasificación, asociación y agrupamiento.

En el módulo de preprocesamiento incorpora una serie de filtros supervisados y no supervisado, a nivel de atributos e instancia. El objetivo es la realización de todas las tareas de limpieza y transformación de datos.
Para nuestro estudio utilizaremos uno de los filtros a nivel de atributos denominado StringtoNominal, que permite convertir un atributo tipo cadena en un tipo nominal.

La incorporación de la herramienta Estadística en Moodle, nos simplificaría bastante el trabajo de recolección de información para un posterior análisis, sin necesidad de incurrir a alguna herramienta externa para tal propósito.
Otra actividad planificada es proceder a efectuar un análisis y sumarización de la información obtenida en cada sesión de los alumnos considerando un determinado periodo de tiempo. Como máximo dos años, que es lo que plantea la herramienta. Es posible que esta información pueda ser migrada a otro formato con el objetivo de facilitar y agilizar su análisis como por ejemplo, a formato de texto u ODS o planilla de

cálculo. El formato seleccionado para este trabajo será planilla de cálculo por lo potencia de las herramientas para el análisis estadístico y representación gráfica de los datos.

Extracción de Conocimiento: esta etapa es la parte central del proceso KDD, el objetivo central es la búsqueda de patrones, en este caso serán los grupos de usuarios con comportamientos de uso de la plataforma virtual educativa Moodle, con características similares.

La tarea principal en esta etapa será la creación de perfiles. Un perfil está compuesto por sesiones de alumnos usuarios similares entre sí, que representan un comportamiento de navegación específico de acuerdo con las páginas que fueron visitadas en cada una de las sesiones.
La construcción de creación de perfiles puede ser usando técnicas explicitas o implícitas.
En nuestro estudio aplicaremos una combinación de ambas, para la generación de perfiles.
Técnica Implícita: se toma como base las acciones pasadas de los usuarios, como hemos detallado en la actividad posterior se tomará información de la actividad extraída de una herramienta estadística de Moodle.
Técnicas Explicitas: consiste en la construcción de un dispositivo para la extracción de conocimiento relacionado con las preferencias de uso de los alumnos.
La combinación de estas técnicas permitirá la obtención de un vector con una ponderación sobre los parámetros considerados.

A continuación se muestra un ejemplo de dispositivo, que es una matriz de calificaciones de los usuarios.

	Navegabilidad	Facilidad de acceso	Facilita Aprendizaje
Usuario 1	5	7	8
Usuario 2	4	8	6
Usuario 3	3	7	7
.....			
Usuario N	6	6	5

Figura 1

En este trabajo se aplicarán técnicas de agrupación, el objetivo es la identificación de grupos de alumnos que han tenido comportamientos similares de navegación y utilización de la plataforma. El algoritmo que usaremos será uno de las más simples y clásico dentro de este tipo de técnicas, que es el K-Means.

El procedimiento a aplicar teniendo en cuenta el algoritmo seleccionado será:

- Se considerarán todas las sesiones de los alumnos, extraídas en la etapa anterior y se seleccionará de manera arbitraria las "n" sesiones que se convertirán en los centros de los "n" grupos.
- De las sesiones restantes, a cada una de ellas se las asignará a un determinado grupo. La pertenencia estará fundamentada por un cálculo. El mismo resulta de la similitud de la sesión del alumno y cada uno de los centros seleccionados en el punto anterior. Será finalmente al grupo de mayor similitud.
- El centroide de cada grupo será recalculado.
- Las sesiones serán nuevamente reasignadas, teniendo en cuenta el establecimiento de los nuevos centroides.
- Los pasos 2,3 y 4 se repetirán, hasta que los nuevos centroides de cada grupo no hayan tenido una variación significativa.

Interpretación y evaluación de resultados: Una vez obtenido los perfiles de los estudiantes, es necesario poder evaluar la calidad o nivel de confianza de los mismos. Las métricas Recall y Precisión utilizadas en el área de recuperación de información, han sido utilizadas para tal fin.
El recall es una medida que corresponde al porcentaje de ítems que son correctamente recomendados, mientras que la precisión mide la calidad promedio de una recomendación individual. La validación de los resultados para cualquier técnica de minería de datos, es fundamental, ya que determinan el nivel de confianza de las inferencias resultantes de este estudio.

3. Resultados

En este trabajo se propone una metodología de minería de uso web aplicada a la enseñanza acompañada y guiada en el uso de plataformas de educación virtual.
Con el resultado de aplicar esta metodología, se espera responder los siguientes interrogantes como:

¿Los alumnos usan la plataforma sólo como herramienta de comunicación con el docente?
¿Cuál es el nivel de acceso a las actividades académicas diseñadas por el docente?
¿Cuál es el nivel de acceso a las actividades de evaluación y autoevaluación, diseñadas por el docente?
¿Los alumnos sólo acceden a los contenidos curriculares de la asignatura, conceptualizando la herramienta como un repositorio de información, sin

tener en cuenta otros recursos como foros, encuestas etc.?

Lograr responder a estos interrogantes, nos permitirá diseñar una alternativa de personalización de plataforma educativa web, que permita tener un nivel más alto de atención y establecer una guía que les facilite a los alumnos poder acceder a información pertinente y hacer uso adecuado de los recursos didácticos fortaleciendo el aprendizaje significativo y autónomo.

4. Discusión

Castaño P. Andrés en su trabajo "Minería de Uso Web para la identificación de patrones", ha realizado investigaciones relacionada con la implementación de reglas de asociación, haciendo énfasis en las reglas difusas.
Las conclusiones obtenidas son que las reglas de asociación clásicas son recomendadas para el caso en que los repositorios de datos contengan por lo general atributos de naturaleza booleana, sin embargo cuando se encuentran atributos ya sean numéricos o nominales que es bastante frecuente, las reglas de asociación son altamente recomendables.
El trabajo con reglas difusas se ha demostrado que se pueden alcanzar mejores representaciones, para obtener patrones de comportamiento de los usuarios.

Fredy Andrés Novoa, Sandra Patricia Tucarruncho Tucarruncho y Arturo Tucarruncho Tucarruncho, en su trabajo "Extracción de Perfiles basada en agrupamiento genético para recomendación de contenido", ellos han considerado para la creación de perfiles técnicas de agrupamiento. Los algoritmos considerados para el estudio han sido KMeans y el algoritmo de Agrupamiento genético.

Este trabajo ha concluido que los resultados arrojados por el algoritmo de agrupamiento genético son más precisos que el algoritmo KMeans.
La fundamentación de estos resultados se debe a que el algoritmo encuentra de forma automática el número apropiado de grupos y los construye basándose en una fuerte similitud entre sesiones y no, como en el caso de k-Means, basándose en el centro.

Si bien este trabajo plantea como propuesta la implementación del algoritmo KMeans, para la creación de perfiles se atenderá cual es el nivel de confianza ofrecido por el modelo, contrastando con otra técnica de agrupamiento que no sea algoritmo de agrupamiento genético para verificar el nivel de precisión.

5. Conclusión

En este trabajo uno de los objetivos ha sido presentar un caso práctico en el que se aplica una minería de uso aplicada al contexto educativo.
Para ello hemos planteado un modelo de obtención de reglas mediante un análisis de preprocesamiento de archivos que están incorporados en la plataforma virtual Moodle.

De los datos contenidos en los archivos se ha realizado una tarea de filtrado con la finalidad de seleccionar un conjunto de atributos que sean los más significativos para aportar una mejor propuesta de solución a la problemática planteada en este trabajo.

La técnicas seleccionada en este trabajo para la generación de perfiles es agrupamiento más específicamente el algoritmo K-Means.

Si bien esta propuesta ha seleccionado este algoritmo uno de los más clásicos, sería un buen punto de partida para futuras investigaciones cual es el grado de precisión del modelo obtenido.

Una alternativa posible es el algoritmo adaptativo, que es un algoritmo heurístico de agrupamiento se puede adaptar a la problemática a resolver, ya que en este caso no se conoce previamente el numero de clases o categorías del problema.

6. Referencias

[1] Enrique García Salcines, Cristóbal Romero Morales, Sebastián Ventura Soto y Carlos de Castro Lozano, "Sistema recomendador colaborativo usando minería de datos distribuida para la mejora continua de cursos e-learning", IEEE-RITA Vol. 3, Núm. 1, Mayo 2008

[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande y Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Paterns from Web Data", SIGKDD Explorations, ACM. Enero 2000.

[3] Antonio González Torres, "Minería web y personalización: Revisión bibliográfica y propuesta de un marco de referencia".

[4] Sandra Patricia Tocarruncho Tocarruncho, Fredy Andrés Aponte Novoa, Arturo Tocarruncho Tocarruncho, "Extracción de perfiles basado en agrupamiento genético para recomendación de contenido", Conferencia IADIS Ibero-Americana WWW/Internet, 2007.

[5] Baeza-Yates, Ricardo, "Excavando en la web", En: El profesional de la información, 2004, enero-febrero, v. 13, n. 1, pp. 4-10.

[6] Cecchini Rocío Luján, , "Minería de la Web: Crawling the Web".

[7] Thorsten Joachims. Universidad Dormunt. Departamento de Informática, Dayne Freitag. Universidad Carnegie Mellon. Departamento de Ciencias de Computación, Tom Michell. Universidad Carnegie Mellon. Departamento de Ciencias de Computación, "Web Watcher: A tour guide for the World Wide Web".

[8] Lic. Sady C. Fuentes Reyes e Ing. Marina Ruiz Lobaina, "Minería Web: un recurso insoslayable para el profesional de la información", VI Jornada Bibliotecaria del IDICT, 17 y 18 de julio de 2007 en el Capitolio Nacional, La Habana, Cuba.

[9] Blaya Juan Botía. Departamento de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia, "Introducción al uso de la Minería Web: Data Mining sobre ficheros log".

[10] Bamshad Mobasher, DePaul University, "Web Usage Mining and Personalization".

[11] Mike Perko, Oren Etzioni, Department of Computer Science. University of Washington. Seatle, "Adaptive Web Sites: an AI Challenge".

[12] Mestras Juan Pavón, Departamento de Sistemas Informáticos y Programación, **"**Personalización de los servicios en la web", Universidad Complutense Madrid, 2001

Datos de Contacto

Cynthia Lorena Corso. Entidad: Universidad Tecnológica Nacional. Facultad Regional Córdoba. Código Postal: 5012.
Email: cynthia@bbs.frc.utn.edu.ar