Integrando diferentes técnicas de Data Mining en procesos de Web Usage Mining

Luca Cernuzzi
Universidad Católica "Nuestra Señora de la Asunción"
Departamento de Ingeniería Electrónica e Informática
Asunción - Paraguay, C.C. 1683, Fax: +595 21 310587
lcernuzz@uca.edu.py

and

María Liz Molas
Universidad Católica "Nuestra Señora de la Asunción"
Departamento de Ingeniería Electrónica e Informática
Asunción - Paraguay, C.C. 1683, Fax: +595 21 310587
lizmolas@telesurf.com.py

Abstract

Web Usage Mining focuses on techniques to search for patterns in the user behaviour when navigating the Web. This work presents a methodological proposal for the integrated application of different Data Mining techniques, in the KDD general process framework, in order to do Web Usage Mining. The methodological proposal is applied to a case study, trying to describe the users' behaviour of a Web portal. Moreover, the study briefly presents the more relevant results obtained during the analysis.
Keywords: KDD, Web Mining, Web Usage Mining, Association Rules, Clustering

Resumen

Web Usage Mining, se basa en las técnicas para buscar patrones en el comportamiento de los usuarios cuando navegan en la Web. En el presente trabajo se presenta una propuesta metodológica para la aplicación integrada de diferentes técnicas de Data Mining, dentro del marco general del proceso de KDD, para realizar Web Usage Mining. Dicha propuesta metodológica es aplicada a un caso de estudio, para intentar describir el comportamiento de los usuarios de un portal Web. También se presentan los resultados más significativos obtenidos durante el análisis.
Palabras claves: KDD, Web Mining, Web Usage Mining, Reglas de Asociación, Clustering

1. Introducción

El crecimiento explosivo de la Web en los años recientes la ha convertido en una gran fuente de datos disponibles on-line. Entre otras características, la Web no provee a sus usuarios páginas con un diseño estándar, es heterogénea en su contenido tanto en relación con la información disponible como a la calidad. En el ambiente Web existe una variedad de datos que pueden ser estructurados, semi estructurados y no estructurados. Además el volumen de datos que se manejan diariamente en los servidores web es muy grande La Web puede ser vista como una colección no estructurada de páginas e hiperlinks, las páginas son accedidas por una gran variedad de personas con diferentes conocimientos e intereses Esto implica una mayor dificultad en inferir conocimiento con respecto a las bases de datos convencionales. Las páginas son generalmente diseñadas con HTML, con una estructura limitada que dificulta el análisis del contenido que realizan las herramientas automáticas. Todos estos hechos denotan la dificultad del usuario para encontrar información relevante en la Web [2,10].

Knowledge Discovery (KDD) y Data Mining son disciplinas de búsqueda que involucran el estudio de técnicas para buscar patrones en grandes colecciones de datos [6].

La aplicación de las técnicas de Data Mining a la Web, llamada Web Data Mining o más sintéticamente Web Mining, es definida como el estudio de las técnicas de Data Mining que automáticamente descubren y extraen información desde la Web [5].

En este contexto, la técnica de WebUsage Mining, intenta descubrir y extraer patrones de uso o comportamiento a partir de datos de la exploración o la navegación (tal como los registros del los archivos log de acceso a los grandes repositorios de datos Web) [3].

El presente trabajo presenta una propuesta metodológica para la realización de Web Usage Mining y presenta un caso de estudio cuyo objetivo es obtener la descripción del comportamiento del usuario durante la visita al sitio web, base importante para la toma de decisiones de diseño del sitio, en términos de su contenido y estructura, también útil para el desarrollo de políticas de Web caching, transmisiones de red y distribución de los datos.

La metodología propuesta para la realización de Web Usage Mining dentro del marco más general de KDD integra dos técnicas de Data Mining (Reglas de Asociación y Clustering).

Como base para el caso de estudio, se utilizaron los archivos log del servidor web de registros de navegaciones para miles de individuos o usuarios en el Portal de Rieder Internet (un Internet Service Provider local).

El resto del trabajo se estructura de la siguiente forma. En la sección 2 se presentan brevemente los conceptos principales para contextualizar y caracterizar las técnicas de Web Mining. En la sección 3 se presenta una propuesta metodológica para la aplicación de dichas técnicas. En la sección 4 una aplicación a un caso de estudio. En la siguiente sección, el análisis de los resultados obtenidos y finalmente se trazan algunas conclusiones y se perfilan posibles trabajos futuros.

2. Web Mining

En los últimos años ha tomado un creciente interés la disciplina de *Knowledge discovery in databases* – KDD, que consiste en el proceso no trivial de identificar patrones validos, novedosos, potencialmente útiles y entendibles, en los datos [6].

El proceso de KDD es interactivo e iterativo, y envuelve numerosos pasos donde muchas decisiones son tomadas por el usuario. La mayoría de los trabajos que se han realizado sobre KDD se centran en el paso de Data Mining. que consiste en la aplicación de algoritmos específicos que, bajo algunas limitaciones aceptables de eficiencia computacional, produce una enumeración particular de patrones [6].

Cabe mencionar que las técnicas de Data Mining han sido usadas para una variedad de tareas o aplicaciones, sin embargo, desde un punto de vista global, Fayyad [6] propone dos categorías de problemas generales, la *predicción* y la *descripción*. La Predicción se basa en algunas variables o campos de la Base de Datos para predecir valores desconocidos o futuros de otras variables de interés. La Descripción, en cambio, se centra en encontrar patrones interpretables por el ser humano, a partir de la descripción de los datos. Para ambos podemos utilizar alguna de las siguientes tareas básicas: clasificación, regresión, clustering, condensación o resumen, modelado de dependencias, análisis de enlaces y análisis secuencial.

Web Data Mining (o simplemente Web Mining) ha sido definida como la aplicación de las técnicas de Data Mining a datos web [2]. En este sentido, en términos amplios Web Mining puede ser definida como el descubrimiento y el análisis de información útil desde la World Wide Web [5].

El objetivo primario del Web Mining es descubrir patrones interesantes en los accesos a varias páginas del espacio web asociado a un servidor particular [12].

El análisis de los datos de acceso del servidor puede proveer información significante y útil que soporten las decisiones de negocios en el comercio electrónico, que ayuden a incrementar el rendimiento, reestructurar el sitio web para mejorar su efectividad y clasificar a los clientes en grupos de interés para dirigir la publicidad o información en general [5,10].

2.1 Web Usage Mining

En la disciplina de Web Mining recubre particular interés la disciplina que se enfoca el análisis de la información de las visitas a distintas páginas Web en orden a extraer patrones de uso, eso es Web Usage Mining.

Cuando los usuarios de la Web interactúan con un sitio, los datos registrados de su comportamiento son típicamente almacenados en los archivos log del servidor web. Estos archivos pueden contener información sobre la experiencia del usuario en el sitio. El tamaño promedio de los archivos puede ser de varios megabytes diarios, por esto son necesarias técnicas y herramientas que faciliten el análisis de su contenido [2].
El análisis de cómo los usuarios acceden a un sitio es critico para determinar la eficacia de las estrategias de marketing y la optimización de la estructura lógica del sitio web. Debido a numerosas características exclusivas del modelo cliente servidor en la Web, incluyendo diferencias entre la topología física de los repositorios y los caminos de acceso de los usuarios, además de la dificultad en identificar a los usuarios en sesiones o transacciones, es necesario desarrollar un nuevo framework para implementar el proceso de minería [5].

3. Propuesta metodológica para Web Usage Mining

El proceso de KDD típicamente involucra una serie de pasos (varios autores reconocen 9 pasos básicos) organizados en un proceso iterativo de 6 etapas. Si bien, en términos generales esto puede resultar interesante y cubre la mayor parte de los posibles métodos para descubrir conocimiento en bases de datos, consideramos que el proceso aplicado al caso de conocimiento inherente a usuarios que navegan en la Web puede ser simplificado. En la Figura 1 se presenta el sistema de Web Usage Mining adaptado al marco más general del proceso de KDD. Además, para el paso inherente a la búsqueda de patrones de comportamiento, es decir, para la aplicación de técnicas de Data Mining, se considera que a diferencia de la mayoría de los enfoque típicamente adoptados de centrarse en una técnica particular, posiblemente la más adecuada al problema en estudio, puede resultar muy útil la integración de distintas técnicas complementarias. Esta misma observación puede ser valida para la etapa de análisis de resultados, en la cual distintas técnicas podrían confirmar los resultados obtenidos o bien ayudar un análisis más detallado en ciertos aspectos abriendo espacio para nuevos patrones.

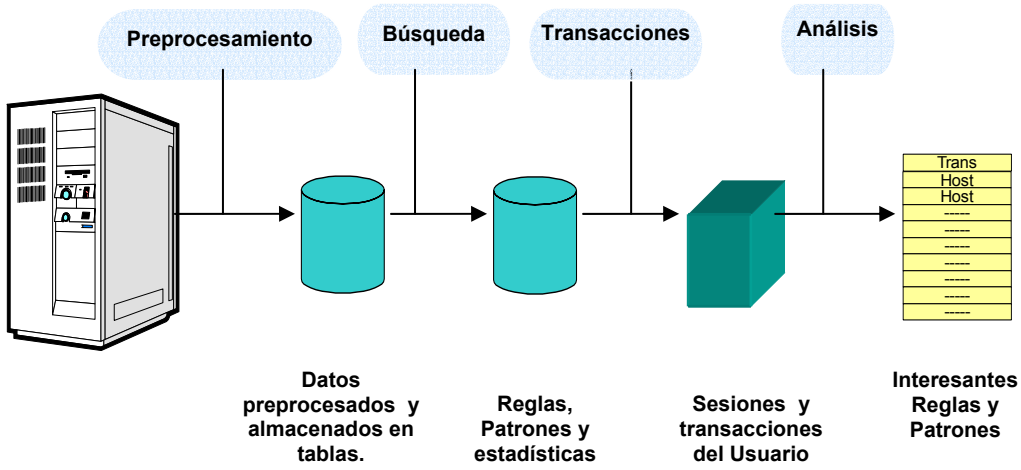

Figura 1 Propuesta metodológica para Web Usage Mining

En el presente trabajo, se presenta la aplicación de la propuesta metodológica a un caso de estudio en el que se modela una colección de sesiones de usuarios a partir de los archivos log de acceso al servidor Web. Para el efecto, se insertan los datos en bruto en tablas relaciónales para facilitar el preprocesamiento y transformándolos en transacciones para luego aplicar en forma integrada dos técnicas estándar de Data Mining ("Reglas de Asociación" y "Clustering"). La aplicación de dos técnicas permite ir validando los resultados de una con la otra y así descubrir nuevos patrones que pueden ser mejor detectables a través de otra técnica. El análisis de los resultados, se realiza mediante la integración de mecanismos de SQL y la técnica de "Visualización" para corroborar los patrones encontrados o descubrir nuevos. Para efectivizar el trabajo de Web Usage Mining, surge la necesidad de contar con un sistema software cuyo objetivo principal sea el preprocesamiento de los accesos log y la identificación de las transacciones de los usuarios, para luego permitir la aplicación de los algoritmos de Data Mining para el descubrimiento de los patrones de comportamiento. También se necesita la creación de una base de datos para almacenar en forma estructurada los accesos log y las transacciones generadas, y como soporte para la aplicación de los algoritmos de Data Mining.
Para el diseño y desarrollo del sistema y de la base de datos y la implementación del proceso de KDD se ha decidido adoptar un enfoque ingenieristico y una metodología orientada a objetos, eso es, el Proceso Unificado Racional [8,9].
Los usuarios que interactúan con el sistema, son el cliente o usuario web quien navega por las páginas del Portal, el Servidor web que registra todos los archivos y páginas solicitados por el usuario web, el analista quien ejecuta los diferentes pasos del Proceso de KDD y el experto que analiza los resultados del Proceso de KDD.

4. Aplicando Técnicas de Web Usage Mining: un caso de estudio

Para poder aplicar y validar las técnicas de Data Mining en general, y en particular aquellas de Web Usage Mining que constituyen el enfoque del presente trabajo, se necesita un gran volumen de datos textuales, y un sitio con una estructura de navegación no trivial. Para el efecto se ha utilizado como caso de estudio el sitio de Rieder Internet (uno de los proveedores de Internet local), y los archivos log de acceso almacenados en su servidor web.

Rieder Internet (http://www.rieder.net.py) ofrece, entre otros servicios, noticias locales e internacionales, asistencia a sus clientes y además soporta otros sub sitios con dominio propio. El sitio cuenta con una variedad de información distribuida en diferentes secciones, tales como mujer, deportes, niños, jóvenes, negocios, tecnología y diversión.

Las técnicas de Data Mining pueden proveer de información cualitativa, en forma de patrones de comportamiento del usuario, para medir la efectividad del sitio, para verificar los supuestos que tiene el contenido de las páginas y para ofrecer un mejor servicio en forma de personalización a sus usuarios.

En este trabajo se ha utilizado un sub conjunto de los archivos log de acceso del servidor web (de abril a julio del 2002), trabajando en forma coordinada con el experto, quien conoce en detalle la estructura del sitio web y los objetivos de la empresa Rieder Internet, en el análisis de los datos y los resultados obtenidos en los diferentes pasos del Proceso de KDD. En la etapa de Data Mining se pretende analizar transacciones de los usuarios para intentar descubrir patrones de comportamiento, es decir describir el comportamiento del usuario cuando navega por el sitio web.

A fin de alcanzar nuestros objetivos hemos seleccionado 2 de las 5 técnicas más conocidas de Data Mining propuestas por distintos autores [5,14]. Las técnicas "Path analysis", que busca los caminos más visitados en el sitio web, y "Sequential patterns**",** que busca las transacciones ordenadas en el tiempo, donde la presencia de un conjunto de elementos es seguida por otro, no están directamente relacionadas con nuestros objetivos por eso desechamos ambas técnicas. También obviamos la técnica de "Classification rules" porque exige una tipificación previa de los usuarios, que no se dispone actualmente. Con la aplicación de la técnica de "Reglas de Asociación" buscamos las asociaciones y correlaciones entre las referencias o accesos a los archivos disponibles en el servidor web. Además, con la técnica de "Clustering" aplicada a las sesiones se intenta inferir los perfiles de usuarios, validar los patrones encontrados y descubrir nuevos.

Como han sugerido Massegkia y Cicchetti [10] las "Reglas de Asociación" pueden ser vistas como la relación entre ciertos hechos almacenados en la base de datos. Los hechos considerados pueden ser simples características o comportamientos observados en los individuos. Dos hechos pueden considerarse relacionados si ocurren para el mismo individuo. Se puede decir que una relación determinada no es relevante si ésta es observada en pocos individuos; por el contrario, si es frecuente puede considerarse un conocimiento importante para la toma de decisiones para quien intenta inferir una descripción general a partir de casos particulares.

Por otro lado la técnica de "Clustering" es la división de los datos en grupos similares de objetos, donde cada grupo es llamado cluster y contiene todos los objetos que son similares entre si y distintos a los objetos de otros grupos. La representación de los datos por medio de pocos cluster necesariamente implica la perdida de los detalles finos, pero ofrece simplificación. Esta técnica representa muchos objetos de datos por medio de pocos cluster, y por tanto, los datos son modelados a través de los cluster [1,3].

Según lo observado en la literatura de Web Usage Mining normalmente se aplican técnicas de Data Mining independientemente, en cambio en nuestro trabajo aplicamos en forma integrada dos técnicas ("Reglas de Asociación" y "Clustering"), para validar los resultados una con la otra y descubrir nuevos patrones que pueden ser mejor detectables a través de otra técnica. Cabe destacar que la primera técnica nos permitió determinar cuales eran las páginas más utilizadas tanto para ingresar como para salir del sitio, mientras la segunda fue de suma ayuda para determinar el comportamiento en general de los usuarios del portal cuando navegan a través del sitio.

4.1 Tareas de Preprocesamiento

Se realizaron varias tareas de preprocesamiento previas a la aplicación de los algoritmos de Data Mining, sobre los datos colectados en los archivos log.

Durante la tarea de **limpieza** se han verificado los sufijos del nombre del URL accedido, para identificar los elementos irrelevantes. Por ejemplo, todos los registros con extensión GIF, JPEG, JPG, y MAP pueden ser removidos por que corresponden a imágenes.

Luego de la identificación de la extensión de los archivos accedidos por los usuarios (10.983.362 registros), se ha procedido a eliminar de la base de datos aquellos que fueran indicadas como irrelevantes por el experto. Así se obtuvo un total de 828.463 registros, que fueron sometidos a su vez a otro proceso de filtrado, identificando todos los que pertenecían al Portal de Rieder. Para ello se ha utilizado la topología o estructura de páginas proveída por el Administrador del sitio web, debido a que el servidor web de Rieder también publica páginas de otras compañías clientes y registra en los archivos log los accesos a estas páginas. Realizados todos los filtros necesarios sobre los datos, se identificaron 481.915 registros validos.

A continuación se ha realizado la tarea de **identificación de los usuarios**, que permitió la identificación de todos los IP que accedieron al servidor web, por un total de 11.833 IP diferentes.

Para los archivos log que pertenecen a largos periodos de tiempo, es muy probable que el usuario haya visitado el sitio web más de una vez. El objetivo de la tarea de **identificación de las sesiones** es dividir los accesos a páginas de cada usuario en sesiones individuales. Un método simple para la identificación de la sesión es tomar todas las páginas accedidas por un usuario identificado en el archivo log y según la heurística utilizada. En nuestro caso se han agrupado los accesos en sesiones de 30 minutos como máximo entre la primera y la última página accedida. Luego se procedió a determinar la longitud del tiempo de acceso de cada página referenciada por los usuarios, mediante el calculo de la diferencia entre la hora de la siguiente referencia y la referencia actual. Para la última referencia de la sesión el tiempo se asume que es 0.
Como resultado de esta tarea fueron identificados 126.664 Sesiones de usuarios.

4.2 Identificación de transacciones

Antes de aplicar las técnicas de minería de datos, las secuencias de páginas referenciadas por los usuarios deben ser agrupadas en unidades lógicas que representen las transacciones web o sesiones de usuarios. Una sesión de usuario es el conjunto de todas las páginas referenciadas por un usuario determinado durante una simple visita al sitio. La transacción difiere de una sesión de usuario en que el tamaño de la transacción puede ser una simple página referenciada o todas las páginas referenciadas en una sesión, dependiendo del criterio utilizado para identificar las transacciones [5].
Varios autores manifiestan que los usuarios tratan a cada página a la que acceden de dos maneras: llaman referencias a *páginas auxiliares* (o *de navegación*) a aquellas utilizadas para encontrar los links a los datos, y las *páginas de contenido*, donde el usuario se detiene porque encuentra información que le resulta interesante. El campo de fecha/hora de acceso y el URL registrados en el log han sido utilizados para clasificar cada acceso como referencia de navegación o contenido para el usuario correspondiente.
Usando los conceptos de referencias a páginas de navegación y de contenido, se proponen dos maneras de definir las transacciones [1,5,12].
El primer método define una transacción como aquella que contiene todas las referencias auxiliares e incluye como última una referencia de contenido de un usuario y sesión determinados. La aplicación de las técnicas de minería de datos sobre este tipo de transacción, da como resultado los caminos más comunes de navegación para acceder a una página de contenido.
El segundo método define una transacción como aquella que contiene todas las referencias de contenido para un usuario y sesión dados, descartando las referencias auxiliares. La aplicación de las técnicas de minería de datos sobre las transacciones solamente de contenido, puede dar como resultado asociaciones entre las páginas de contenido.

5. Análisis de Patrones

De acuerdo a algunos autores [11] la cuestión fundamental en la aplicación de los algoritmos de Data Mining, es antes que nada conocer la utilidad de los algoritmos para la clase de problemas a ser considerados. En otras palabras, se debe conocer bien antes de empezar el proceso de KDD el problema particular P, con las características Cj del tipo de problema o tarea, para determinar los algoritmos específicos de Data Mining Ai que tengan mejor desempeño, para resolver el problema P.
También se requiere que el usuario interactué sobre varios pasos del KDD, especialmente cuando los resultados no son muy buenos, en términos de precisión o entendimiento (claridad) de las reglas generadas para el modelo [11].
El análisis de los Patrones es el último paso en el Proceso de Web Usage Mining. El objetivo del análisis de los patrones es filtrar las reglas o patrones que no sean interesantes dentro del conjunto encontrado en la fase de descubrimiento de patrones.
La metodología exacta de análisis está gobernada por la aplicación para la que se realiza el Web Mining. La forma más común de análisis de patrones consiste en mecanismos de consultas tal como *SQL*. Otro método es a través de cubos de datos en orden a desarrollar operaciones *OLAP* (On-Line Analytical Processing). Además, las técnicas de *Visualización*, tal como patrones gráficos o asignación de colores a diferentes valores, pueden dar una visión general de los patrones o tendencias en los datos [5,14].

5.1 Resultados obtenidos con la Técnica "Reglas de Asociación"

Analizando las reglas obtenidas con la técnica de "Reglas de Asociación", aplicada sobre los distintos archivos generados según la longitud de accesos se pueden observar ciertas similitudes entre las reglas. Los archivos fueron generados agrupando las transacciones con la misma longitud o número de accesos a páginas (referencias a páginas).
Cabe aclarar algunos términos relacionados con las Reglas de Asociación. Así el **soporte (s)** de una regla se refiere a el número de ocurrencias del conjunto de elementos en la base de datos de transacciones, mientras que la **confianza (α)** es el porcentaje de transacciones que contienen a todos los elementos que componen la regla [12].
En el contexto de Web Usage Mining, con la aplicación de esta técnica se intenta descubrir todas las asociaciones y correlaciones entre las páginas accedidas dentro de las sesiones de los usuarios, ignorando las que tienen un soporte

pobre, es decir las que no aparecen en un número suficiente de transacciones [12,14]. Para nuestro caso de estudio, cada transacción está compuesta por las páginas accedidas por un usuario dentro de una misma sesión, contemplando solo las referencias de contenido.

Archivos	Nro	Regla	Confianza
Transacciones de 3 páginas accedidas	R4	PAG1=/ PAG2=/ 422 ==> PAG3=/ 348	0.82
	R8	PAG2=usados.php 68 ==> PAG3=usados.php 49	0.72
	R13	PAG1=viewthread.php PAG3=viewthread.php 210 ==> PAG2=viewthread.php 129	0.61
	R20	**PAG2=post.php 151 ==> PAG1=viewthread.php 78**	**0.52**
	R42	PAG2=forumdisplay.php 163 ==> PAG1=viewthread.php 56	0.34
	R75	PAG1=/ 707 ==> PAG3=viewthread.php 63	0.09
Transacciones de 4 páginas accedidass	R1	PAG2=usados.php PAG4=usados.php 21 ==> PAG3=usados.php 19	0.9
	R4	PAG2=publicaraviso.php 15 ==> PAG3=publicaraviso.php 12	0.8
	R6	PAG1=u2u.php PAG2=u2u.php PAG4=u2u.php 22 ==> PAG3=u2u.php 17	0.77
	R66	**PAG2=post.php 121 ==> PAG1=viewthread.php 62**	**0.51**
	R100	PAG1=post.php 113 ==> PAG2=forumdisplay.php 18	0.16
	R192	PAG2=forumdisplay.php 127 ==> PAG1=viewthread.php 33	0.26

Tabla 1 Reglas extraídas sobre archivos que agrupan transacciones de 3 y 4 páginas accedidas

Por ejemplo se observan un par de reglas en la tabla 1, que indican que el 50% de los clientes que acceden a /post.php, accedieron antes a /viewthread.php. Esto indica que alguna información en /viewthread.php direcciona a los clientes a post.php; cabe destacar que ambas páginas pertenecen al sub sitio "La Cueva".
También se puede notar que para algunas reglas el porcentaje de confianza es elevado, mientras su representatividad es baja. Por ejemplo una regla nos indica que el 100% de los clientes que ingresan en la primera y quinta referencia en /u2u.php, también acceden a la misma página en la cuarta referencia de la transacción. Pero el número de transacciones que cumple con está regla representa el 6% de la muestra de datos, según el análisis realizado mediante consultas SQL a la base de datos.
Podemos decir que de todas las transacciones analizadas, un total de 9989, el 43% inicia la sesión en la página principal del Portal '/', el 20% en la página viewthread.php y el 7.5% en la página forumdisplay.php, el resto se distribuye uniformemente entre todas las demás páginas del portal. Esto nos indica que un buen porcentaje de transacciones se inicia a través de la página principal del Portal de Rieder, donde se encuentran los link a todos los demás sub sitios, y que otro grupo importante inicia su visita al Portal a través del sub sitio La Cueva. En la tabla 2 se presenta la relación de las transacciones con la última página accedida antes de finalizar la visita, observada en las transacciones agrupadas según el tamaño de la transacción, es decir la cantidad de páginas accedidas.

Tamaño trans.	Última Página	Porcentaje(%)
Dos páginas accedidas 4212 registros	/	32.2
	viewthread.php	16.2
	Index	10
Tres páginas accedidas 2194 registros	/	23.3
	viewthread.php	19.7
	Index	13.3
Cuatro páginas accedidas 1285 registros	viewthread.php	26.5
	/	12.5
	Index	12
	Post.php	10.5
Cinco páginas accedidas 935 registros	viewthread.php	28.3
	Index	13
	/	7.3
Seis páginas accedidas 559 registros	viewthread.php	27.5
	Index	13
	/	8.5
Siete páginas accedidas 400 registros	viewthread.php	25.7
	Post.php	16
	/	7.7
Ocho páginas accedidas 237 registros	viewthread.php	26
	Post.php	21
	/	8
Nueve páginas accedidas 114 registros	viewthread.php	35
	Post.php	12
	/	3.5
Diez páginas accedidas 53 registros	Post.php	24
	viewthread.php	22
	/	4

Tabla 2. Última página accedida en las transacciones de tamaño de 2 a 10 páginas.

Se han seleccionado las transacciones de 2 a 10 referencias o accesos, ya que representan aproximadamente el 99% de las transacciones que posibilitan la aplicación la técnica "Reglas de asociación". Observando la tabla 2 se puede notar que las páginas "/" y "viewthread.php" han sido utilizadas como última página en cerca del 50% de las transacciones de longitud 2, donde el 32 % pertenece a la página "/" principal del Portal. En las transacciones de mayor longitud el porcentaje de transacciones que utilizan la página principal disminuye regularmente, en cambio el porcentaje de transacciones que utiliza la página "viewthread.php" aumenta. Lo que significa que los usuarios que acceden a varias páginas dentro de una transacción, terminan su visita en páginas del sub sitio "La cueva".

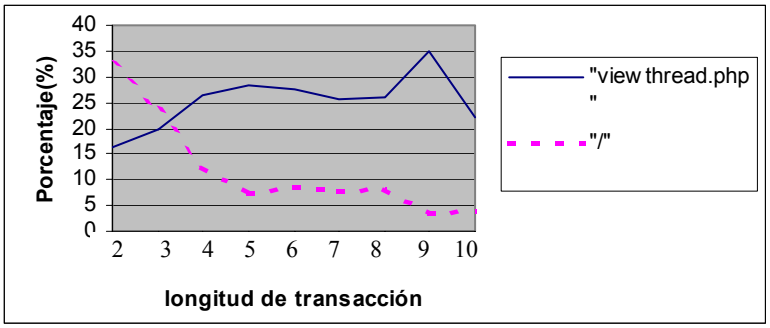

Figura 2. Última página accedida en las transacciones de tamaño de 2 a 10 páginas.

En la figura 2 se presenta la variación del porcentaje de accesos a las páginas que mayor porcentaje alcanzaron, estas son la página "/" principal del Portal y "viewthread.php".

5.2 Resultados obtenidos con la Técnica "Clustering"

Analizando los cluster obtenidos con la técnica de "Clustering", aplicada sobre los distintos archivos generados según la longitud de accesos de las sesiones, se pueden observar ciertas similitudes entre ellos.
Cabe mencionar que la estructura del sitio tiene 170 páginas diferentes, distribuidas en 6 sub sitios. Para facilitar la aplicación de la técnica de "Clustering" y el posterior análisis de los resultados, en la generación de los archivos de sesiones se reemplazaron los valores de los nombres de las páginas accedidas por los sub sitios a los que pertenecen, estos son: rieder, lacueva, elladrillo, newsletter, riederusaros, webmouse y principal para la página del Portal.
En los resultados pertenecientes a los archivos de sesiones de 2 y 3 accesos a páginas, que en conjunto representan el 70% de la muestra analizada, se puede observar que el cluster más resaltante para ambos archivos presenta un porcentaje superior al 50%, y corresponde a accesos a la página principal. Este resultado es importante porque confirma la regla encontrada en el análisis realizado con la técnica de "Reglas de Asociación" donde se descubrió que el 30% de transacciones se iniciaban en la página principal. En la Tabla 3 se presentan los cluster mencionados.

Archivo de sesiones de 2 accesos(26322)	
Cluster 1:	14327 (61%)
principal, principal	
Archivo de sesiones de 3 accesos(11075)	
Cluster 0:	5681 (51%)
principal, principal, principal	

Tabla 3. Cluster mayoritario encontrado en los archivos de sesiones de 2 y 3 accesos

Archivo de sesiones de 2 accesos(26322)	
Cluster 0:	7996 (34%)
principal, rieder	
Archivo de sesiones de 3 accesos(11075)	
Cluster 8:	2487 (22%)
principal, rieder, rieder	
Archivo de sesiones de 4 accesos(5950)	
Cluster 3:	998 (17%)
principal, principal, rieder, principal	
Archivo de sesiones de 5 accesos(3495)	
Cluster 1:	1019 (29%)
principal rieder rieder rieder rieder	

Tabla 4. Cluster encontrado en los archivos de sesiones de 2 a 5 accesos

Otra coincidencia interesante encontrada en los resultados de la aplicación de la técnica en los archivos de sesiones de 2 a 5 accesos a páginas consiste en la existencia de un cluster con un porcentaje entre 15% a 30 % de accesos tanto a la página principal como al sub sitio Rieder. Este resultado no fue observado con la técnica de "Reglas de Asociación", pero si fue descubierto como un comportamiento en un grupo importante de sesiones de usuarios. Cabe mencionar que este grupo de archivos en conjunto representa el 87% de la muestra analizada. En la Tabla 4 se presentan los cluster que contienen accesos a la página principal y al sub sitio rieder.
En los resultados de los demás archivos se ha notado una distribución cada vez más uniforme a medida que se incrementa el número de accesos por sesión.
Surgen 3 cluster importantes en el análisis de los resultados obtenidos a partir de los archivos de 6 a 10 accesos, el primero es el conjunto que incluye solo accesos a la página principal, el segundo contiene accesos al sub sitio de "La cueva" y el tercero es el conjunto de accesos al sub sitio de "Rieder". A continuación, en la tabla 4, se presentan estos resultados en detalle.
Otra coincidencia observada en los resultados de los archivos de 9 y de 10 accesos es el cluster que incluye en mayor parte accesos al sub sitio "La cueva", este comportamiento confirma los resultados obtenidos con la técnica de "Regla de Asociación", donde notamos que para los archivos de transacciones de tamaño 9 y 10 accesos, el último a página de las transacciones corresponde al sub sitio "La cueva".

Archivo de sesiones de 6 accesos (2300)	
Cluster 2:	757 (33%)
principal rieder rieder rieder rieder rieder	
Cluster 0:	656 (29%)
principal principal principal principal principal principal	
Cluster 1:	384 (17%)
lacueva lacueva lacueva lacueva lacueva lacueva	
Archivo de sesiones de 7 accesos (1606)	
Cluster 0:	511 (32%)
principal principal principal principal principal principal principal	
Cluster 9:	365 (23%)
principal rieder rieder rieder rieder rieder rieder	
Cluster 2:	296(18%)
principal lacueva lacueva lacueva lacueva lacueva lacueva	
Archivo de sesiones de 8 accesos (1153)	
Cluster 2:	284 (25%)
principal principal principal principal principal principal principal principal	
Cluster 0:	267 (23%)
principal lacueva lacueva lacueva lacueva lacueva lacueva lacueva	
Cluster 5:	225 (20%)
principal rieder rieder rieder rieder rieder rieder rieder	
Archivo de sesiones de 9 accesos (955)	
Cluster 8:	244 (26%)
principal lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva	
Cluster 3:	184 (19%)
principal rieder rieder rieder rieder rieder rieder rieder rieder	
Cluster 5:	146 (15%)
principal principal principal principal principal principal principal principal principal	
Cluster 1:	126 (13%)
principal principal principal principal rieder rieder principal rieder rieder	
Cluster 6:	107 (11%)
principal rieder newsletter newsletter newsletter newsletter newsletter newsletter newsletter	
Archivo de sesiones de 10 accesos (772)	
Cluster 9:	128 (17%)
principal rieder rieder rieder rieder rieder rieder rieder rieder rieder	
Cluster 8:	127 (16%)
principal principal principal principal principal principal principal principal principal principal	
Cluster 2:	115 (15%)
lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva	
Cluster 5:	114 (15%)
principal lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva lacueva	

Tabla 5. Cluster encontrado en los archivos de sesiones de 6 a 10 accesos

6. Conclusión y trabajos futuros

En este trabajo se ha investigado y analizado la utilidad de técnicas de Data Mining aplicadas a la información Web [2,5,14], dentro del marco de Knowledge Discovery [6,11,13].

Se han presentado los pasos del proceso de Web Usage Mining y sus resultados sobre un conjunto de datos reales considerando como caso de estudio el portal de Rieder Internet, uno de los proveedores de Internet en nuestro país. Cabe destacar que, para la fase de minería de datos se adoptaron diferentes técnicas del área de Data Mining sobre el dominio especifico. En particular se utilizaron las técnicas de "Reglas de Asociación" y "Clustering".

Finalmente, con la colaboración del Experto se analizaron los patrones y reglas resultantes de la aplicación de las técnicas "Reglas de Asociación" y "Clustering" mediante el uso de los mecanismos de SQL y la técnica de "Visualización".

Entre los aportes más significativos se pueden citar:

- Propuesta de una metodología adoptando técnicas especificas de Web Usage Mining en el marco más general existente para KDD.
- Según lo observado en la literatura de Web Usage Mining normalmente se aplican técnicas de Data Mining independientemente, en cambio en nuestro trabajo aplicamos en forma integrada dos técnicas ("Reglas de Asociación" y "Clustering"), para validar los resultados una con la otra y descubrir nuevos patrones que pueden ser mejor detectables a través de otra técnica.
- El análisis de los resultados, mediante la integración de mecanismos de SQL y la técnica de "Visualización" para corroborar los patrones encontrados o descubrir nuevos.
- Además, se han presentado algunos resultados particulares obtenidos en el caso de estudio. Esto es, el descubrimiento de patrones como ser la existencia de tres grupos de usuarios con comportamientos resaltantes: el primer grupo accede solo a la página principal del Portal de Rieder, la mayoría en sesiones de corta duración (menor a 5 accesos a páginas); el segundo grupo ingresa al Portal a través de la página principal pero luego requiere páginas del sub sitio "Rieder"; y el tercer grupo accede a páginas del sub sitio "La cueva", este último grupo utiliza generalmente sesiones de larga duración (mayor a 5 accesos a páginas).

Algunas posibles áreas de interés que requieren futuras investigaciones y desarrollo son:

- Reestructuración del sitio, personalización y desarrollo de políticas de Web caching.
  En particular, utilizando información sobre las preferencias de los usuarios, que han sido obtenidas mediante técnicas de Descripción, el resultado del Web Usage Mining y técnicas de Data Mining para Predicción, se puede lograr la anticipación a la necesidad del usuario Web y proveerle información personalizada cuando visita las paginas del Portal.
- Desarrollo de Herramientas inteligentes, que puedan asistir en la interpretación de los conocimientos descubiertos, sabiendo que el resultado de los algoritmos de Data Mining generalmente no tiene un formato adecuado para el análisis de los usuarios. Claramente, este tipo de herramientas necesita conocimiento especifico sobre el dominio del problema en particular. En Web Mining, podría ser adecuado el uso de agentes inteligentes desarrollados sobre la base de los patrones de accesos descubiertos, la topología del sitio Web y ciertas heurísticas.

Bibliografía

[1]. Berkhin Pavel, "Survey of Clustering Data Mining Techniques",Accrue Software, (2002) http://citeseer.nj.nec.com/berkhin02survey.html

[2]. Cabral de Moura Borges José Luís, "A Data Mining Model to Capture User Web Navigation Patterns" . Department Of Computer Science, University College London (2000). http://www.fe.up.pt/~jlborges/publications/BorgesPhDthesis.ps.gz

[3]. Cadez Igor, Heckerman David, "Visualization of Navigation Patterns on a Web Site Using Model Based Clustering" (2000) http://citeseer.nj.nec.com/292620.html

[4]. Cooley R., Mobasher B. and Srivastava J., "Data Preparation for Mining World Wide Web Browsing Patterns" (1999). http://maya.cs.depaul.edu/~classes/ect584/papers/cms-kais.pdf

[5]. Cooley R., Mobasher B., "Web Mining: Information and Pattern Discovery on the World Wide Web" (1998) http://maya.cs.depaul.edu/~mobasher/classes/ds575/papers/Webmining.pdf

[6]. Fayyad Usama, Piatersky-Shapiro Gegory, Padhraic Smyth, "From Data Mining To Knowledge Discovery", Jet Propulsion Laboratory California Institute Of Technology. AAAI Press / The MIT Press. Menlo Park, California, USA. 1996, In ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, ISBN 0-262-56097-6, pag 1-34 (1996) http://www.aaai.org/Press/Books/Fayyad/Fayyad-preface.pdf

[7]. Hay Birgit, Wets Greert and Vanhoof Koen, "Clustering navigation patterns on a website using a Sequence Alignment Method" (2000) http://citeseer.nj.nec.com/451556.html

[8]. Jacobson Ivar, Booch Grady, Rumbaugh James,"El Proceso Unificado de desarrollo de Software". Rational Software Corporation. Pearson Educación, Madrid. ISBN 84-7829-036-2, pag 4-8 (2000)
[9]. Jacobson Ivar, Booch Grady, Rumbaugh James, "El Lenguaje Unificado de Modelado". Rational Software Corporation. Addison Wesley Iberoamericana, Madrid ISBN 84-7829-028-1. (1999)
[10]. Massegkia Florent, Ponclet Pascal, Cicchetti Rosine, "An efficient algorithm for Web usage mining", Universite de Versilles, Francia. (2000)
http://citeseer.nj.nec.com/399609.html
[11]. Meneses Claudio J. , Grinstein Georges G.,"Categorization And Evaluation Of Data Mining Techniques", Departamento de Ing. de Sistemas y Computación, Universidad Católica Del Norte, Antofagasta-Chile . Proc. of Intl. Conf. on Data Mining, Sept. 1998, In DATA MINING (Ebecken, N.F.F., editor), ISBN 1853126772 pag. 53-80 (1998)
[12]. Mobasher B. and Srivastava J., "Web Mining: Pattern Discovey from World Wide Web Transactions" (1996)
http://citeseer.nj.nec.com/mobasher96web.html
[13]. Rezende S.O., Oliveira R.B.T, "Visualization for Knowledge Discovery in Database", Departament of Computer Science and Statistics, Institute of Mathematical and Computer Sciences, University of Sao Paulo.1996, In DATA MINING (Ebecken, N.F.F., editor), ISBN 1853126772 pag. 81-95 (1996)
[14]. Srivastava Jaideep, Cooley Robert, Deshpande Mukund, Tan Pang-Ning, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", Department of Computer Science and Engineering. University of Minnesota(2000)
http://www.cs.umn.edu/research/websift/papers/sigkdd00.ps