Descubrimiento de Conocimiento en Portales Institucionales

Luis Alberto Olguin, Alum. Sebastián Lobo, Mag. Raúl Klenzi, Mag. Alejandra Malberti
Instituto de Informática (IdeI) / Departamento Informática (DI) / Facultad de Ciencias Exactas Físicas y Naturales (FCEFN) / Universidad Nacional de San Juan (UNSJ)
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", San Juan
{lolguinunsj, lobo.sebastian7, rauloscarklenzi, amalberti} @gmail.com

Resumen

El presente trabajo trata el proceso de descubrimiento de conocimiento a partir del conjunto de datos generados por el Log File de un servidor web institucional. Se describen los métodos y técnicas utilizadas aplicando la herramienta de software libre RapidMiner (RM) 5.3.15
Los resultados obtenidos por estos estudios buscan apoyar la toma de decisiones en la organización respecto al re-diseño de la página institucional y estrategias de fidelización a partir de los patrones de navegación descubiertos. Como caso de aplicación se analizan logs del sitio web de la Biblioteca Franklin de la ciudad de San Juan.

Palabras clave: Data Mining, Web Mining, Log´s de Servidor, Sitio Web.

Contexto

La línea de investigación se enmarca en el proyecto trianual 2011-2013 "**MINERÍA DE DATOS (MD) EN LA DETERMINACIÓN DE PATRONES DE USO Y PERFILES DE USUARIO**" código 21/E889 que se desarrolló en el ámbito de la Facultad de Ciencias Exactas, Físicas y Naturales de la Universidad Nacional de San Juan (FCEFN-UNSJ), aprobado por el Consejo de Investigaciones Científicas Técnicas y de Creación Artística (CICITCA), financiado por la propia Universidad y ajustado a evaluación externa.

Introducción

Desde hace muchos años, los centros de documentación y particularmente la biblioteca Franklin de San Juan realizan procesos de registro de sus actividades tanto administrativas, como las asociadas a las transacciones de su área de circulación y préstamo. Estas actividades de registro pretenden tener exactitud a fin de no generar conflictos, es así que se detallan fechas, horas, documentos, etc. Con el paso del tiempo estos "datos" han tenido un crecimiento importante. Sin embargo en la institución solo son utilizados como herramientas de consulta.
Similar situación se da en la página web institucional. Allí, día a día se generan cientos de entradas que son almacenadas en los *log server file* algunos de los cuales simplemente se pierden por falta de procesos de resguardo. Esta acumulación de datos ha generado inquietud por descubrir qué relación es posible encontrar entre ellos.
La MD es el campo en el que se aplican técnicas para el análisis de conjuntos de datos observacionales, con el objeto de encontrar relaciones insospechadas y resumirlas para que sean interpretables y útiles para el propietario de los datos.[1]
Los recursos informativos que un centro de información coloca en la web forman

parte de la colección que mantiene la biblioteca por tanto es de sumo interés analizar si la información es útil y si es fácilmente ubicable por el visitante web.
Si la información publicada no se usa, el tiempo y recurso humano involucrado en su creación se transforma en una mala inversión para la institución.
Los datos transaccionales del servidor si son recuperados, transformados y analizados desde la óptica de la MD, permiten descubrir tanto el comportamiento de los usuarios frente al sitio web como las posibles relaciones entre los contenidos visitados, que hasta el momento podrían ser desconocidas.[2]
La aplicación en bibliotecas de herramientas de software libre, como GreenStone, KOHA entre otros, es un tema de especial interés para la institución ya que forma parte de la política de migración de sus sistemas a plataformas web.
De igual manera una herramienta de software libre como RM, de aplicación en el área del aprendizaje automático, extracción de conocimiento y referenciada por la comunidad científica, es una excelente plataforma para realizar la manipulación, selección y procesamiento de los datos de logs y la posterior aplicación de técnicas de minería de datos para descubrir asociaciones o correlaciones entre los ítems accedidos. En este caso se aplicaron reglas de asociación para determinar la relación entre los movimientos de un visitante en el sitio web de la institución.[2][3][4]

Líneas de investigación y desarrollo

El proyecto "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuario" y que dio marco de contención a la presente propuesta expiró el 31 de diciembre pasado. Por ello, y dando continuidad a las tareas de investigación, el grupo de investigadores elevó una nueva propuesta para el bienio 2014-2015 que se encuentra actualmente en etapa de evaluación. En esta oportunidad se pretende continuar con las líneas de investigación anteriores extendiéndolas al tratamiento de grandes datos, haciendo uso de hardware paralelo (CPU multinúcleos, Unidades de Procesamiento Gráfico GPUs, cluster de computadoras) y de herramientas de software libre que soporten dicho hardware, que tengan implementaciones de algoritmos de minería de datos paralelos o permitan la escritura de nuevas propuestas. Una de estas herramientas de software libre, RM bajo licencia AGPL, es la que el grupo ha utilizado para el preporcesamiento y posterior análisis de logs del servidor de la biblioteca Franklin.
En este caso se aplicaron las tareas de: 1) ingreso de datos, ajustando los parámetros de los diferentes módulos de RM al formato de los datos de entrada 2) preprocesamiento por medio de filtrados, ajustes y adaptación de formatos de datos, necesarios para en el paso 3) descubrir patrones de visita frecuentes mediante reglas de asociación.

Desarrollo

1) Etapa de recolección y lectura de la fuente de datos a través del operador Read Server Log-Figura 1.

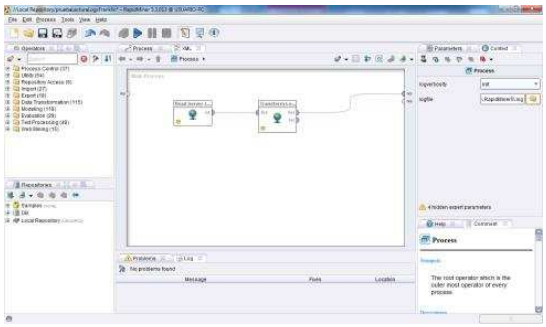

Figura 1- Lectura de Fuente de Datos

En esta lectura, se procesan todos los datos que generan las actividades realizadas por un usuario, en particular dentro de nuestro sitio web. Así, si un usuario accede a una página con nombre "buscadordelibros.html", se añadirá una entrada al log correspondiente, con la información generada por ese usuario al momento de acceder a la página en cuestión. Este es un ejemplo típico de un registro de log:

223.10.215.126 - - [08/Feb/2002:13:45:29 +0100] "GET /index.html HTTP/1.0" 200 42898 "-" "Mozilla/4.0 (compatible; MSIE 5.0; Windows 98)"

En él se pueden observar diferentes tipos de datos, como ser fecha y hora de acceso, url o dirección web accedida, protocolo de conexión, tipo de navegador, entre otras.

2) Etapa de Pre-procesamiento, donde se eliminan aquellos datos redundantes, o que no aporten información significativa. A través del operador GenerateAttribute se crea un nuevo atributo, llamado "*urinew*", que permite hacer el filtrado de aquellos datos innecesarios. A partir de ello el operador Filter Examples detecta datos poco significativos. Luego el operador Select Atribute, captura la salida, filtrando y eliminando todas aquellas entradas marcadas en el paso previo.

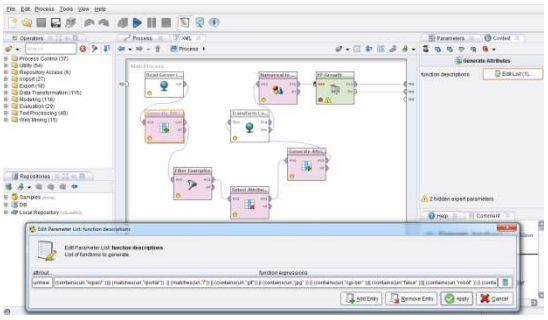

Figura 2- Filtros Aplicados

La Figura 2 presenta los filtros utilizados para eliminar entradas con imágenes (.jpg, .png, .gif, .ico), accesos a estructuras del servidor web (cgi-bin), hojas de estilos (.css), robots para buscadores web, archivos de javascript (.js), y entradas incompletas o no válidas. Además, es necesario realizar la transformación de los datos en sesiones de usuarios, las cuales identifican unívocamente todas las entradas del log. Este proceso se lleva a cabo mediante el operador Transform Log to Session. Por último, solo resta convertir el tipo de datos a una forma binomial (verdadero o falso) a través del operador Numerical to Binomial para dejar los datos listos para ser procesados. Figura 3

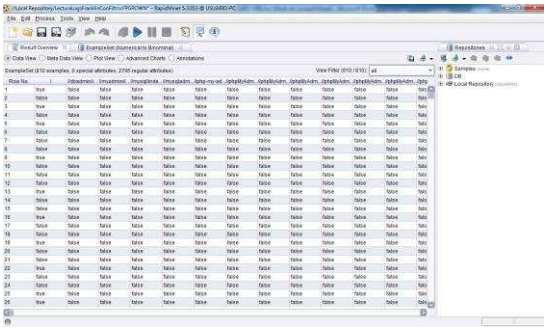

Figura 3- Transformación de entradas en sesiones de usuarios y conversión de datos a binomiales

3) Etapa de descubrimiento de patrones: Una vez que las tareas 1) y 2) se han realizado, los datos de la web están listos para descubrir en ellos nuevos patrones. Estos patrones pueden encontrarse a partir de la aplicación de análisis estadístico estándar, algoritmos de agrupamiento, reglas de asociación, algoritmos de clasificación, y descubrimiento de patrones secuenciales.

En esta oportunidad se hizo uso de reglas de asociación. En este modelo, en el que cada transacción consta de un conjunto de ítems, el problema consiste en encontrar relaciones entre ítems, llamadas reglas de asociación, a partir de la presencia frecuente de varios ítems dentro de esas transacciones.

Formalmente se considera [5][6][7][8]:

D={ d1, d2,..., dm} un conjunto de ítems,

T={ t1, t2,..., tn} un conjunto de transacciones, donde cada transacción ti es un conjunto de ítems tal que ti $\subseteq$ D 1<=i<=n.

La implicación X $\Rightarrow$ Y es una Regla de Asociación donde X$\subset$D ,Y$\subset$D , X$\cap$Y= ϕ, los conjuntos X e Y son mutuamente excluyentes, y X $\cup$ Y $\subseteq$ ti , esto es, el conjunto de ítems formado por aquellos que corresponden al antecedente o al consecuente de la regla de asociación, debe estar contenido o ser igual a alguna de las transacciones pertenecientes a T.

La regla X $\Rightarrow$Y tiene soporte *s* en el conjunto de transacciones T, 0<=*s*<=1, si s% de las transacciones de T contienen tanto a X como a Y.

En el contexto de WebMining este problema tiende a descubrir la correlación entre los accesos de los clientes a varios ítems disponibles en el servidor.[9][10]

Cada transacción está compuesta por un conjunto de URL accedidas por el cliente en una visita al servidor.

Para poder obtener las reglas de asociación de manera correcta, se usa el operador FP-Growth provisto por la herramienta. El mismo calcula de manera eficiente todos los conjuntos de elementos frecuentes. Cabe destacar que todos los atributos de entrada deben ser binominales, de no ser así los datos no serían procesados.

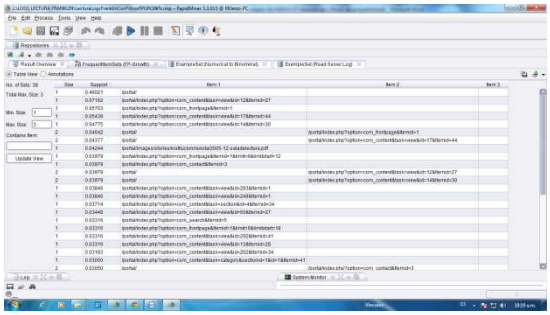

Figura 4- Obtención de reglas de asociación

De los resultados obtenidos y el valor de soporte entregado por la herramienta -Figura 4-, podemos observar, por ejemplo, que más del 46% (0,46021) de los usuarios acceden a la página principal, mientras que el 4,37% (0,04377) de los usuarios acceden primero a la página principal, y luego ingresan a la sección de servicios de la Biblioteca (Itemid=44).

Resultados y Objetivos

Resultados:

Al momento de la elaboración de este documento y a través del uso de reglas de asociación, se obtuvo una buena aproximación de las relaciones existentes entre las diferentes páginas accedidas por los usuarios que visitan el sitio.

Objetivos:

Se propone a modo de ampliación:

- Completar este estudio mediante la utilización de otras herramientas de software que, conjuntamente con RM, permitan descubrir el landing page (página de aterrizaje).

- Agregar o mejorar la publicidad de los diferentes eventos llevados a cabo por la Biblioteca, en aquellas páginas que son accedidas con mayor frecuencia.

Esto ayudaría a dar respuesta a los siguientes interrogantes:

Es posible disminuir la tasa de rebote en el sitio mejorando su diseño? en otras palabras, ¿La estructura, contenidos y apariencia del sitio es óptima e invita a los usuarios a permanecer en él, o presenta una interfaz poco amigable y difícil de usar y/o comprender?.
¿Es posible determinar la exit page (página de salida) para evaluar, en ese momento, la satisfacción del visitante?,

Formación de RRHH

La ejecución de las tareas proyectadas incidirá directamente en una formación más profunda de los integrantes del equipo de investigación en las tecnologías de Data Mining, Webmining y Bibliomining.
Además, se prevé la generación de Trabajos Finales de Grado para las carreras Licenciatura en Sistemas de Información, y Licenciatura en Ciencias de la Computación.

Referencias

[1] Hand, David; Mannila, Heikki; Smith, Padhraic, (2001) "*Principles of Data Mining"*.The MIT Press.
[2] Liu Bing (2007) "*Web Data Mining. Exploring Hyperlinks,Contents, and Usage Data*". Springer-Verlag.
[4] Markov, Zdravko; Larose, Daniel T.(2007). "*Data mining in the webs"*.Uncovering Patterns in Web Content, Structure, and Usage. 218 Seiten, Hardcover -Praktikerbuch-ISBN-10: 0-471-66655-6. ISBN-13: 978-0-471-66655-4 - John Wiley & SonsInc Publication.
[4] Chakrabarti Soumen (2003). "*Mining. The Web Discovering Knowledge From Hypertext Data*" Morgan Kaufmann .
[5] Hand, David; Mannila, Heikki; Smith, Padhraic (2001). "*Principles of Data Mining"*.The MIT Press.
[6] Kantardzic, M. (2003) *"Data Mining: Concepts, Models, Methods, and Algorithms"*.John Wiley & Sons.
[7] Klenzi R. (2008).*"Aplicación de minería de datos a la gestión bibliotecaria"*. Tesis de Maestría. Universidad Nacional de la Matanza.
[8] Larose, Daniel T. (2006). *"Data mining methods and models"*.Department of Mathematical Sciences.Central Connecticut State University.John Wiley & Sons, Inc Publication.
[9] Parr-Rud, Olivia(2001)**.***"Data Mining Cookbook.Modeling Data for Marketing, Risk, and Customer Relationship Management"*.Published by John Wiley & Sons, Inc.
[10] Wu, C.H. (2003). "*Data mining applied to material acquisition budget allocation for libraries: design and development"*. Expert Systems with applications.

Sitios en internet:
11) Rapid-I.http://rapid-i.com/api/rapidminer-5.1/com/rapidminer/tools. 2011
12)http://www.mkt-sapiens.com.ar/docs/Guia-de-Web-Analytics_-_Resultics-2010.pdf
13)http://www.thinkepi.net/