ANALITICA WEB EN CENTROS DE INFORMACION

Prog. Luis Alberto Olguín; Mag, Raúl Oscar Klenzi
Instituto de Informática – Departamento de Informática
Facultad de Ciencias Exactas, Físicas y Naturales
Universidad Nacional de San Juan
lolguin@iinfo.unsj.edu.ar; rauloscarklenzi@gmail.com;

Resumen

En la actualidad muchos centros de información se encuentran en la carrera de "tener presencia en la web", pero estas acciones, en su mayoría, no son resultado de un análisis previo del "usuario-objetivo" a quien dirigir la oferta web.
Analizar "que hace un visitante en nuestra sede web" permite determinar patrones de comportamiento a fin de optimizar entre otras, la disposición de los enlaces en el sitio web. Desde la incorporación de computadoras en los centros de información se han intentado recuperar estadísticas de uso, para el caso de la web, la primer alternativa son los registros de transacciones del servidor. Las técnicas de analítica web, web mining complementan esta primera aproximación estadística, ofreciendo la posibilidad de análisis más profundos tales como determinar el comportamiento del visitante en el sitio.
Se abordan conceptos generales acerca de web mining y web analytics y se presentan algunas herramientas gratuitas para comenzar la tarea de análisis web, tal es el caso de Awstats y Google Analytics.

Palabras claves: Web Mining, Web Analytics, Transactions log, Transaction Analysis, Usage Statics

Contexto

Este trabajo se enmarca en las actividades previstas en los proyectos "Minería de datos en determinación de perfiles de uso y perfiles de usuarios" (Código 21/E-889) y "Colecciones Digitales para la Facultad de Exactas " (Código 21/E-903), ejecutados por el Instituto Informática y el Departamento de Informática de FCEFN.
Estos proyectos fueron aprobados durante la convocatoria para la ejecución de proyectos bianuales realizada por CICTCA-UNSJ (2011-2012).

Introducción

"En internet, el recuento de las páginas de procedencia de las visitas que llegan a una sede *web* importa y mucho" [1]
Una creciente cantidad de Centros de Información sanjuaninos está volcando información y servicios en la Web mediante el uso de manejadores de contenidos y con la publicación de sus catálogos en línea (OPAC)
Lo que no han realizado estas Instituciones, al menos de manera organizada, es analizar el tráfico web, es decir responder a consultas del tipo "cómo llegan los usuarios a sus páginas?", "se trata de tráfico directo o a través de resultados de motores de búsqueda?", "utiliza enlaces web desde otros sitios?".
En este tipo de organizaciones un factor importante es determinar "la fidelización del usuario", es decir aquel visitante frecuente a la oferta web de la institución. Otro factor de análisis es conocer si la visita viene dirigida desde un enlace del mismo portal de la Institución, como puede ser un link desde alguna de las aulas virtuales (si posee este tipo de servicios).
"Bajo el chaparrón digital que cae, conviene saber de dónde vienen las gotas de agua y nos interesa saber hacia dónde y

cómo orientan las bibliotecas el paraguas (o el embudo para recoger el agua!).
Por qué analizar estadísticas de uso web?: "Without a strategy, the Web's too big". [2]
Los centros de información deben tener claro cuál es su objetivo en la web y luego de socializar este objetivo entre el propio personal, armar una estrategia para su puesta on line.
Logrado esto, hay que plantearse si conocidos los resultados de un análisis web, se toman las acciones que corrigen falencias o no. Para qué analizar el tráfico web si no se actuará correctamente?.
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 ubicada por el visitante web.
"Si la información publicada no se usa, el tiempo y recurso humano involucrado en la creación se transforma en una mala inversión de recursos para la institución"[3]

Minería Web (Web Mining –WM-)
La extracción de información implícita de los datos, previamente desconocida y potencialmente útil es el objetivo buscado por la minería de datos (Data Mining – DM-).
WM se refiere al proceso global de descubrir información o conocimiento potencialmente útil y previamente desconocido a partir de datos de la Web.

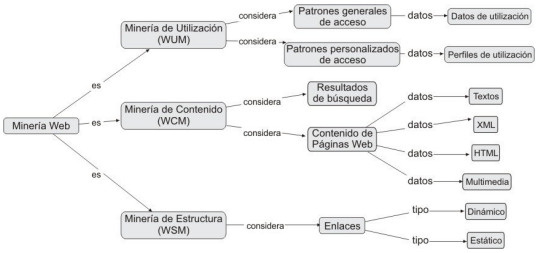

Figura 1 Clasificación MW según Juan Carlos Dürsteler [4]

Dentro del campo de la minería web, esta se puede dividir en tres grandes áreas: minería web de contenido, minería web de estructura y minería web de uso web.
"La minería de uso Web tiene dos enfoques principales; uno es la búsqueda de patrones de acceso general, que analiza el tráfico para entender los patrones de acceso y comportamiento habitual de los usuarios y sus tendencias con el fin de reestructurar el sitio Web ubicando los contenidos de forma más accesible o para ubicar y dirigir a los usuarios de la Web hacia lugares relevantes e importantes para ellos; la segunda tendencia es la búsqueda para personalizar el uso, en la que se analizan las tendencias individuales de cada visitante de la Web para personalizar o adaptar dinámicamente la información del sitio Web, su estructura o recursos a cada visitante según el patrón de acceso que exhiba"[5].
Para llevar a cabo el proceso de MW de uso web, se establecen cuatro fases:

1. Recolección de datos: Consiste en la recuperación automática de la información relevante para su posterior procesamiento.
2. Procesamiento de los datos. Una vez recuperados los documentos, se ordenan y se preparan para la próxima etapa; se utilizan herramientas para obtener información valiosa en forma automática.
3. Descubrimiento de patrones. Existen múltiples técnicas, aplicables al descubrimiento de patrones. Entre ellas, agrupamiento y clasificación, para el establecimiento de reglas de asociación y el hallazgo de secuencias frecuentes.
4. Análisis de patrones. Comprende su interpretación y validación.

Analítica Web
La analítica web es el conjunto de herramientas, técnicas, métodos que permiten recolectar, analizar y reportar (muy importante!) datos de Internet con el

objetivo de entender cómo es el comportamiento de los usuarios de la web en relación a nuestro sitio web.

El uso de analítica web en las bibliotecas permite ir un poco más allá de las repuestas que se encuentran al usar el análisis de logs de transacciones del servidor y contestar preguntas más estratégicas del tipo "es usable mi sitio?", "la inversión en publicidad on line reporta visitas fidelizadas?", "cuáles son los hábitos de navegación del visitante en nuestro sitio web?".

Conocer estos datos requiere de herramientas tecnológicas apropiadas y de analistas capaces de interpretarlas y extraer información útil para la institución. Estas conclusiones deben impactar positivamente para realizar mejoras en las estrategias "on line" de las bibliotecas.

"Web Analytics es acerca del negocio y no del sitio web. Web Analytics es una disciplina y no solo una herramienta. Web Analytics debería ser parte de la cultura de cualquier organización que realice algún tipo de acción en línea. Web Analytics es acerca de equivocarse, aprender de los errores y corregirlos" [6]

Líneas de investigación y desarrollo

Herramientas para analítica web

"La combinación de herramientas adecuadas y análisis de los datos llevarán al éxito de la estrategia online " [7]

Las primeras herramientas para la analítica web datan de los años noventa y están centradas en el uso de los registros de transacciones del servidor (log server).

Estos archivos almacenan todos los eventos que se producen durante el normal funcionamiento del servicio y en general responden a la norma Common Log format.

El formato de una línea del archivo de logs se ve en la Figura 2.

```
190.245.226.119 - - [21/May/2011:14:33:29 -0300] "GET /biblioteca/opac/buscar.html HTTP/1.1" 200 903 "http://www.bibliotecafranklin.org.ar/portal/" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.17) Gecko/20110420 Firefox/3.6.17 ( .NET CLR 3.5.30729) FBSMTWB"
```

Figura 2: Archivos de Logs

El formato del registro permite determinar: **IP** del cliente remoto, **ID** del usuario remoto (o guion si no está definido), **ID** del usuario **validado** contra el servidor (o guion si no está definido), **fecha de petición**, **petición enviada** por el cliente (URL y método), **estatus** del resultado, bytes del resultado, **referer** (dirección de donde proviene el cliente), **agente de usuario** (la versión del navegador usado por el cliente).

La lectura de estos archivos permite obtener numerosa información, como ser:

- Número de peticiones (hits)
- Número de peticiones por tipo de archivos
- Direcciones de procedencia (referer)

Actualmente existen numerosos softwares libres (Open Source) que permiten la lectura e interpretación de estos archivos. En general, los resultados son presentados como páginas web, lo que permite publicar esta información como parte de la oferta del sitio web (estadísticas).

"AWStats es una herramienta open source de informes de análisis web, apta para analizar datos de servicios de Internet como un servidor web, streaming, mail y FTP. AWstats analiza los archivos de log del servidor, y con base a ellos produce informes HTML. Los datos son presentados visualmente en informes de tablas y gráficos de barra. Pueden crearse informes estáticos mediante una interfaz de línea de comando, y se pueden obtener informes on-demand a través de un navegador web, gracias a un programa CGI." [8]

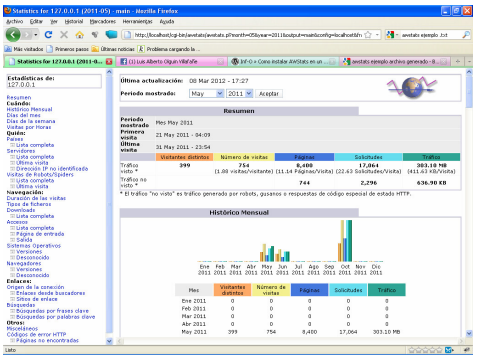

Figura 3 Pantalla principal de AWSTATS – Resumen general

AWSTATS, cuyo entorno se observa en las Figura 3 y 4, permite mostrar, entre otras, la siguiente información:

- Número de visitas y número de visitantes únicos
- Duración de las visitas y últimas visitas Usuarios autenticados y últimos usuarios autenticados
- Días de la semana y horas de mayor tráfico (páginas, hits, KB por cada hora y día de la semana)
- Páginas más vistas, páginas de entrada y salida
- Tipos de archivo
- Navegadores utilizados (páginas, hits, KB por cada usuario, versión, etc.)
- Sistemas Operativos usados Visitas de robots (307 robots detectados)
- Buscadores, palabras clave, frases clave usadas para encontrar tu sitio
- Errores HTTP

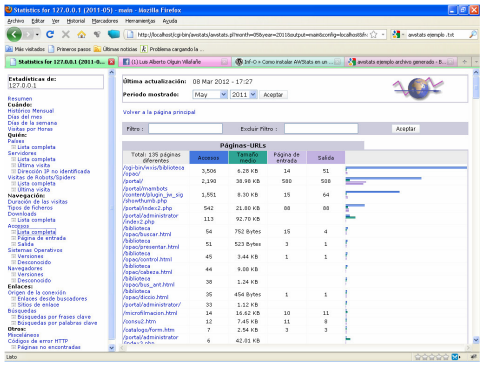

Figura 4: AWSTATS – Detalle de Navegación-Accesos

En la actualidad, es posible ir más allá del análisis de los archivos de transacciones buscando determinar patrones de comportamiento de los visitantes a un sitio web usando herramientas de analítica web, tal es el caso de **Google Analytics** [9] dos de cuyos entornos se observan en la Figuras 6 y 7.

Esta aplicación, gratuita, permite la generación de reportes estadísticos acerca de "número de visitas frecuentes", "páginas visitadas en el sitio", "click-hot" para determinar los links más visitados del sitio, origen del enlace al sitio, "página de salida del sitio", es decir, podemos conocer el comportamiento del usuario de nuestro sitio, desde antes de llegar y hasta que salga.

Para ejecutar su tarea, es necesario que se inserte un "código javascript" en las páginas que se desea analizar. Es justamente este "código de seguimiento" el encargado de recoger los datos y posteriormente enviarlos a una base de datos (en el sitio de google analytics) para posteriormente consultarlo en formato de informe.

Para realizar el análisis, google analytics define algunos parámetros principales [10]

Usuario: Cada objeto que proporciona una cookie de sesión.

Visita: Cada vez que un usuario inicia sesión en nuestro sitio.

Páginas vistas: El número de páginas que accede en cada sesión el usuario.

Porcentaje de rebote: Es un indicador para medir la calidad de las visitas a nuestro sitio.

Sesión: Período de interacción entre el navegador de un usuario y un sitio web concreto, que finaliza cuando se cierra el navegador o al salir de éste es importante mencionar que se considera que una sesión ha finalizado si el usuario ha estado inactivo en el sitio web durante 30 minutos.

```
<script type="text/javascript">
var gaJsHost = (("https:" == document.location.protocol) ? "https://ssl." : "http://www.");
document.write(unescape("%3Cscript src='" + gaJsHost + "google-analytics.com/ga.js' type='text/javascript'%3E%3C/script%3E"));
</script>
<script type="text/javascript">
try {
var pageTracker = _gat._getTracker("UA-10413546-1");
pageTracker._trackPageview();
} catch(err) {}
</script>
```

Figura 5: Código de seguimiento GA para un sitio web

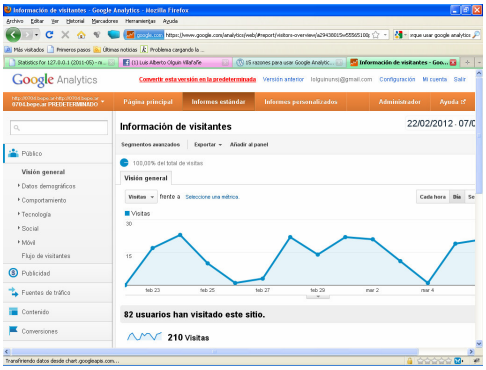

Figura 6: Google Analytics - Visión General

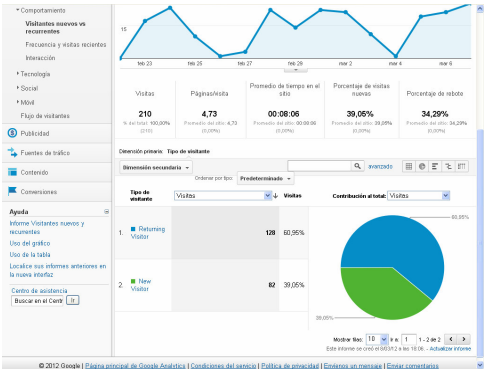

Figura 7: Google analytics – Visitantes nuevos vs. recurrentes

El uso de este tipo de herramientas permite identificar a que áreas geográficas pertenecen los visitantes, determinar las páginas de entrada al sitio y mediante el seguimiento de la navegación, cual es el comportamiento del visitante.

Informes del tipo Bounce Rate (porcentaje de rebote) permite analizar el porcentaje de visitantes que tan solo ven la home page y después dejan el sitio. Determinar "cuan interesante es el sitio" es posible mediante el análisis de permanencia (sesión).

Resultados y Objetivos

Desde los proyectos involucrados, la propuesta es combinar la información proporcionada por las herramientas de software citas anteriormente para tratar de demostrar que la adopción de tecnologías apropiadas, unida a la posibilidad de combinar los resultados obtenidos por estas, permite obtener mejoras en la planificación y puesta en marcha de estrategias web para los centros de información.

Formación de Recursos Humanos

En esta instancia se está dirigiendo un trabajo final de la carrera Licenciatura en Sistemas de Información orientada a técnicas de posicionamiento web, y el conocimiento adquirido se vuelca sistemáticamente en asignaturas de la carrera.

Referencias

[1] http://www.thinkepi.net/
[2] http://www.oclc.org/worldcat/web/default.htm
[3] http://www.haworthpress.com/store/product.asp?sku=J122
[4] http://bibliotecarios.cl/conferencia_2006/C2006_019.pdf
[5] http://vector.ucaldas.edu.co/downloads/Vector4_4.pdf
[6] http://www.mkt-sapiens.com.ar/docs/Guia-de-Web-Analytics_-_Resultics-2010.pdf
[7] http://www.evocaimagen.com/cuadernos/cuadernos2.pdf
[8] http://es.wikipedia.org/wiki/Awstats
[9] http://google.com/analytics
[10] http://recursosweb.unam.mx/pdf/apuntesGA.pdf