Minería del uso de webs

Por **José-Luis Ortega** e **Isidro F. Aguillo**

Resumen: *El análisis del consumo de información en la web es una importante herramienta cibermétrica para conocer no sólo las visitas y visitantes que recibe una sede, sino también los patrones de comportamiento, lo que puede ayudar a diseñar su estructura y contenidos. Se describen los conceptos básicos (ficheros log, sesiones, usuario) y se presentan técnicas manuales y automáticas para el estudio de los ficheros log. Por último se describen los contadores de visitas, ofreciendo un estudio comparativo entre las estadísticas ofrecidas por dos sistemas: Google Analytics y StatCounter.*

Palabras clave: *Cibermetría, Minería web, Análisis de ficheros log, Estadísticas de uso.*

Title: Web usage data mining

Abstract: *The analysis of web usage information is an important cybermetric tool for describing the visits and visitors to a website, and their navigation behaviour, which can help in designing the structure and contents of the webpages. The paper describes the main concepts (log file, session, user) and introduces both manual and automated techniques to study the log files. Finally, the counters and trackers are described, paying special attention to the statistics offered by two systems: Google Analytics and StatCounter.*

Keywords: *Cybermetrics, Web data mining, Log file analysis, Usage statistics.*

Ortega, José-Luis; **Aguillo, Isidro F.** "Minería del uso de webs". *El profesional de la información*, 2009, enero-febrero, v. 18, n. 1, pp. 20-26.

DOI: 10.3145/epi.2009.ene.03

José Luis Ortega Priego *es licenciado en documentación por la Universidad de Granada y actualmente cursa doctorado en documentación en la Universidad Carlos III de Madrid. Forma parte del Laboratorio de Cibermetría del Centro de Ciencias Humanas y Sociales del Consejo Superior de Investigaciones Científicas (CSIC), que investiga en las áreas de la cibermetría, minería web, visualización de información y usabilidad.*

Isidro F. Aguillo *trabaja en el Laboratorio de Cibermetría del Centro de Ciencias Humanas y Sociales del CSIC realizando estudios sobre indicadores web, revistas electrónicas y posicionamiento en motores de búsqueda. Es el editor de la revista-e Cybermetrics y coordina el Webometrics Ranking of World Universities. Es licenciado en zoología –practicando de forma activa la ornitología–, y master en información y documentación por la Universidad Carlos III.*

Análisis de ficheros de transacciones

Desde el surgimiento de la world wide web los ficheros de transacciones (*web log files*) –registros de todas las páginas visitadas por los usuarios que queda en el servidor que aloja la web– han cobrado un gran interés debido a su valor para conocer el uso que se está haciendo de una web. Para el sector comercial muestran información útil para el marketing y las campañas de promoción de servicios y productos comercializados desde la misma (**Gomory** et al., 1999). Desde el punto de vista científico los ficheros de transacciones son una herramienta muy útil para conocer los flujos de información (**Thelwall**, 2001), mejorar el diseño de la web y la estructura de la información (**Spiliopoulou**, 2000). Por último, el mundo de las bibliotecas y de la documentación hace años que se interesa por este tipo de información con la intención de evaluar el uso y la calidad de sus catálogos, además de conocer los hábitos y necesidades de información de sus usuarios (**Peters**, 1993; **Kurth**, 1993).

Análisis de sesiones

Es un tipo de estudio más avanzado y algo más complejo, pero que permite obtener una información de más valor que en el análisis anterior.

Objetivo del análisis de logs

Conocer:

- quiénes nos visitan
- qué necesidades de información tienen
- qué información consumen
- cómo está estructurada nuestra información
- situación de nuestra web en el contexto general de la WWW.

Artículo recibido el 17-12-08
Aceptación definitiva: 07-01-09

Una sesión se define como un periodo continuo de tiempo en el que un usuario está viendo páginas o aplicaciones web dentro de un servidor o dominio. Analizando las sesiones podemos conocer el comportamiento de los usuarios cuando acceden a nuestro sitio web y qué caminos toman para conseguir sus objetivos.

La minería web detecta patrones y localiza información implícita a partir de grandes volúmenes de datos web"

El fin de este examen es conocer la navegabilidad, la organización de los contenidos y los eventuales fallos de diseño para corregirlos en posteriores versiones. Pero para estudiar una sesión debemos individualizar los accesos por cada usuario y conocer todo el recorrido que hace en nuestra web, cosa que tiene sus dificultades. Para solventarlas el análisis de sesiones se ha provisto de técnicas basadas en la minería de datos.

Minería de datos

Llamada en inglés *data mining*, es la extracción de información implícita de los datos, previamente desconocida y potencialmente útil; o la búsqueda de relaciones y patrones globales que existen en las bases de datos (**Klevecz**, 1999). O sea, estamos hablando de una serie de técnicas cuya finalidad es obtener información oculta que está presente en un gran volumen de datos, como puede ser en una base de datos.

La aplicación de esta técnica en el ámbito web ha generado la minería web (*web mining*).

Las técnicas de minería web se han desarrollado en general para poder medir las masivas cantidades de datos e información que se distribuyen por la Red, en concreto en el campo de la cibermetría (*cybermetrics* o *webometrics*) y de la recuperación de información (*information retrieval*). Más específicamente, la minería de datos se ha aplicado a los archivos de transacciones registradas en los servidores web, con la intención de analizar las sesiones que se realizan. La minería de uso web (*web usage mining*) permite individualizar a cada usuario y reconstruir el recorrido completo de su visita.

Minería de uso web

Surge en los noventa con el sistema *Webminer* (**Cooley** et al., 1997, 1999) y se puede definir como el proceso de aplicar técnicas de minería de datos al descubrimiento de patrones de uso a partir de datos web. Con él se pretendía obtener las principales pautas de conducta de los usuarios cuando acceden a una web.

Se estructura en cuatro etapas:

- limpieza de datos
- identificación del usuario
- identificación de la sesión
- reconstrucción de la sesión

Limpieza de datos

En esta fase se elimina todo tipo de accesos que pueden entorpecer el análisis. Debido a que nuestro interés está en las páginas web, eliminaremos los logs correspondientes a los objetos y elementos gráficos que acompañan al contenido de un documento html (ficheros .gif, .jpg, .bmp, etc.). También se eliminarán los scripts y archivos de programación (.js, .cgi, .pl, etc.) que permiten ejecutar funciones pero no aportan contenido en sí.

"No es fácil diferenciar las visitas de los robots, pero al menos se pueden excluir de ciertas partes de la sede con la ayuda del fichero de texto robots.txt"

Por otro lado existen accesos que no son realizados por personas sino por robots o *crawlers* que visitan la web con la intención de indexarla. Es el caso de los robots de los buscadores (*bot*s o *spiders*) como *Google* (*Googlebot*) o *Yahoo! Search* (*Slurp*). Este tipo de accesos es irrelevante a no ser que interese saber qué robots nos visitan y con qué frecuencia. No es fácil diferenciar las visitas de los robots, pero al menos se pueden excluir de ciertas partes de la sede con la ayuda de un fichero de texto llamado "robots.txt", que es leído por los *crawlers* que siguen las normas de buenas prácticas. En este fichero se le dice al robot no solo qué partes de nuestra web queremos que visite y cuáles no, sino la frecuencia de visitas y otros parámetros. Si no queremos ninguna restricción en las visitas de estos programas simplemente no creamos el fichero.

Identificación del usuario

En esta etapa se asocian las páginas de referencia con la misma IP. Esta tarea es complicada por la existencia de cachés locales, cortafuegos corporativos (*firewalls*) y servidores proxy (**Pitkow**, 1997). Para la identificación del usuario se usan los accesos junto con las páginas de referencia y la topología del sitio para construir sesiones de navegación. Los servidores proxy hacen que distintos usuarios tengan la misma IP. Por

ello también se identifica el navegador, su versión y el sistema operativo con el fin de reducir los accesos coincidentes y se puedan detectar usuarios diferentes. Como último paso para separar usuarios idénticos se analizan las propias sesiones de navegación. Por ejemplo, si de una página A no se puede pasar a una página B, porque no existe ningún enlace entre ellas, el acceso a la página B corresponde a un usuario distinto del que accede a la página A.

Identificación de la sesión

El objetivo ahora es dividir los accesos de un mismo usuario en distintas sesiones. La forma más simple es establecer un tiempo límite.

Catledge y **Pitkow** (1995) sugirieron 25 minutos como tiempo máximo, aunque generalmente se estima que media hora entre un acceso y otro es la medida adecuada.

Reconstrucción de la sesión

Puede suceder que se hayan producido accesos que el fichero de transacciones no ha recogido, mermando así la construcción de la sesión. Como se dijo, esto se produce por las copias caché que utilizan los ordenadores o servidores proxy del usuario. La solución es contrastar la página de referencia con la página solicitada y comprobar si existe un enlace entre ellas y así reconstruir aquellos accesos que no han sido computados por el fichero de transacciones. Existen otros procedimientos más complejos, pero en este caso sólo hemos considerado como sesiones a los accesos con las páginas de referencia identificadas.

En el siguiente ejemplo encontramos accesos que el fichero de transacciones no ha recogido:

> 161.111.229.64 - [29/Nov/2007:14:37:26 +0100] "GET /servicios/busquedas_bibliograficas.html HTTP/1.0" 200 11948 "http://www.cindoc.csic.es/servicios/servicios.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"
>
> 161.111.229.64 - [29/Nov/2007:14:37:48 +0100] "GET /servicios/biblioteca_principal.html HTTP/1.0" 200 8322 "http://www.cindoc.csic.es/servicios/servicios.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"
>
> 161.111.229.64 - [29/Nov/2007:14:37:53 +0100] "GET /servicios/revistas_electronicas.html HTTP/1.0" 200 8105 "http://www.cindoc.csic.es/servicios/biblioteca_principal.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"

Nuestra sospecha se fundamenta en el conocimiento de la web, su sistema de navegación y la interconexión de sus páginas. Por ejemplo, el referente del segundo acceso:

[http://www.cindoc.csic.es/servicios/servicios.html]

no coincide con la petición anterior

[/servicios/busquedas_bibliograficas.html]

Como consecuencia pensamos que el usuario ha pasado por otra página distinta. Conociendo la estructura de nuestro sitio web sabemos que necesariamente ha vuelto a la página:

[servicios/servicios.html]

El motivo por el que la solicitud no ha sido registrada es porque el usuario ha utilizado el botón "Atrás" de su navegador y éste ha cargado una copia caché interna, por lo que no ha realizado la petición a nuestro servidor.

Finalmente, la ruta quedaría reconstruida como sigue:

> 161.111.229.64 - [29/Nov/2007:14:37:26 +0100] "GET /servicios/busquedas_bibliograficas.html HTTP/1.0" 200 11948 "http://www.cindoc.csic.es/servicios/servicios.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"
>
> 161.111.229.64 - [29/Nov/2007:14:37:30 +0100] "GET /servicios/servicios.html HTTP/1.0" 200 11948 "http://www.cindoc.csic.es/servicios/busquedas_bibliograficas.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"
>
> 161.111.229.64 - [29/Nov/2007:14:37:48 +0100] "GET /servicios/biblioteca_principal.html HTTP/1.0" 200 8322 "http://www.cindoc.csic.es/servicios/servicios.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"
>
> 161.111.229.64 - [29/Nov/2007:14:37:53 +0100] "GET /servicios/revistas_electronicas.html HTTP/1.0" 200 8105 "http://www.cindoc.csic.es/servicios/biblioteca_principal.html" "Mozilla/3.01 [es] (Win95; I; 16bit)"

Como se ha visto, a través de una serie de pasos se han podido identificar las sesiones solventando los problemas con cachés e IPs. En este ejemplo no existe una identificación explícita del usuario a través del Nombre de usuario y Contraseña, ni tampoco a través de *cookies*. En sedes web donde sí exista esta identificación la etapa de Identificación del usuario es innecesaria.

Algunos de los pasos anteriores se han podido realizar con simples consultas, pero existen otros más complejos para los que es necesario programar scripts o pequeñas rutinas en *Visual Basic*, *Java*, *C+* y otros lenguajes para poder extraer los datos.

Finalmente, en la figura 1 se muestra una representación gráfica multidimensional de las sesiones a partir de varias variables. Para ello se ha contado con el software especializado en minería de datos *3dv8 Enterprise*:

http://www.advfn.com/3dv8/

Se diferencian con colores los accesos producidos a nuestra web desde distintos referentes. La altura expresa la profundidad y longitud de la sesión. De esta manera se encuentran pautas en los accesos, se detectan anomalías y fallos en la navegación en las sesiones, se

sabe qué partes son más visitadas y qué caminos hay que tomar para llegar a ellas, etc.

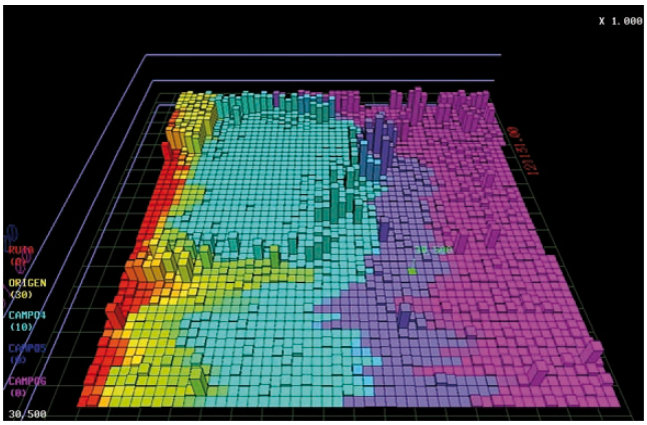


Figura 1. Representación tridimensional de las sesiones mediante 3dv8 Enterprise

Reglas de asociación

A partir de las sesiones pueden extraerse reglas de asociación que permiten relacionar las visitas entre sí y las páginas que son visitadas, lo cual ayuda a detectar patrones o pautas de comportamiento repetibles. Se trata de buscar la relación entre dos páginas visitadas en la misma sesión.

Para ello se definen dos conceptos clave:

Soporte (*sop*): Proporción de visitas que contienen tanto a la página A como a la B. Indica la proporción de casos para los cuales el consecuente es verdadero (lado derecho de la regla).

$$sop\left(A \longrightarrow B\right) = sop\left(A \cup B\right)$$

Si por cada 10 visitas 4 han accedido a la página A y B, se tiene:

Sop = 4/10 = 0,4

El soporte de la relación A y B es 0,4.

Confianza: Proporción de visitas que contienen la premisa y también la conclusión. La confianza entre A y B es la proporción de visitas que entran en la página A y B (soporte de A y B) entre las visitas que sólo acceden a la página A. Si se tenía como soporte 0,4 y el soporte sólo a A es 0,7, la confianza de A y B es 0,4/0,7 = 0,57.

$$conf\left(A \longrightarrow B\right) = \frac{sop\left(A \cap B\right)}{sop\left(A\right)}$$

Si el soporte es suficientemente alto y el tamaño de la base de datos es grande, entonces la confianza es un estimador de la probabilidad de que cualquier transacción futura que contenga la premisa, contenga también la conclusión.

En la figura 2 se ofrece el informe final del software *WebMining log sessionizator 7.0*

http://www.webmining.cl/productos/wlsxp70.asp

Contadores de visitas y estadísticas de uso

El estudio de las visitas se puede realizar de tres formas:

– analizando los ficheros log en el servidor web;

– controlando los accesos desde el cliente;

– utilizando un contador o programa de estadísticas.

El funcionamiento del contador es muy sencillo

El proveedor del servicio de contador proporciona un código característico en formato html que debe añadirse a todas las páginas de la sede que desee controlar. Se puede repetir el mismo código para todas las páginas o bien se pueden utilizar códigos diferentes para secciones o páginas sobre las que se quiere hacer un seguimiento diferencial.

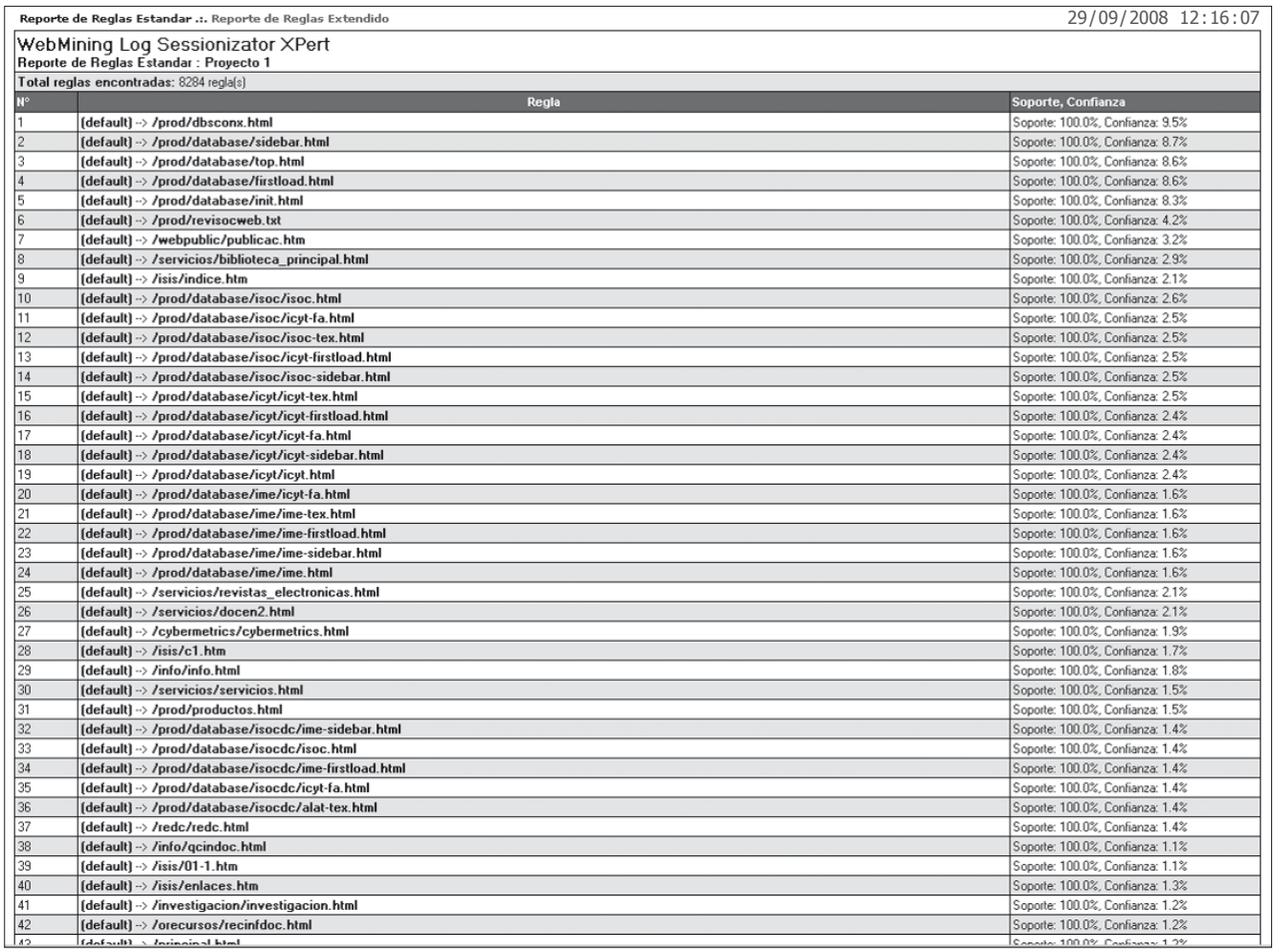
Cada visita es interceptada por el código que remite dicha información al proveedor, el cual la procesa en tiempo real y genera estadísticas, listados, tablas y gráficos al momento. Esa información se publica en página o páginas específicas a las que se accede mediante palabra clave o directamente en abierto.

"La tipología de información que se recoge es muy variada y la oferta de proveedores disponibles muy amplia"

La tipología de información que se recoge es muy variada, y en la tabla 1 se muestra una lista no exhaustiva de variables. *Google Analytics* ofrece un paquete básico de 80 informes más un API que permite configurar adicionalmente otros aspectos, aunque esto requiere conocimientos de programación en *JavaScript*.

La oferta de proveedores disponibles es muy amplia, y en la mayoría de los casos se trata de servicios gratuitos (con una cierta carga de publicidad) o bastante económicos. Hay situaciones intermedias con una oferta básica gratuita y otra *premium*, con más opciones, de pago. La tabla 2 ofrece un listado de proveedores. Aunque todos ofrecen similares prestaciones, las interfaces gráficas varían considerablemente y es aconsejable una revisión previa antes de seleccionar uno cualquiera.

La utilización de este tipo de sistemas tiene indudables ventajas (coste muy bajo, gratuito en muchos



Reporte de Reglas Estandar .:. Reporte de Reglas Extendido 29/09/2008 12:16:07

WebMining Log Sessionizator XPert
Reporte de Reglas Estandar : Proyecto 1

Total reglas encontradas: 8284 regla(s)

N°	Regla	Soporte, Confianza
1	{default} -> /prod/dbsconx.html	Soporte: 100.0%, Confianza: 9.5%
2	{default} -> /prod/database/sidebar.html	Soporte: 100.0%, Confianza: 8.7%
3	{default} -> /prod/database/top.html	Soporte: 100.0%, Confianza: 8.6%
4	{default} -> /prod/database/firstload.html	Soporte: 100.0%, Confianza: 8.6%
5	{default} -> /prod/database/init.html	Soporte: 100.0%, Confianza: 8.3%
6	{default} -> /prod/revisocweb.txt	Soporte: 100.0%, Confianza: 4.2%
7	{default} -> /webpublic/publicac.htm	Soporte: 100.0%, Confianza: 3.2%
8	{default} -> /servicios/biblioteca_principal.html	Soporte: 100.0%, Confianza: 2.9%
9	{default} -> /isis/indice.htm	Soporte: 100.0%, Confianza: 2.1%
10	{default} -> /prod/database/isoc/isoc.html	Soporte: 100.0%, Confianza: 2.6%
11	{default} -> /prod/database/isoc/icyt-fa.html	Soporte: 100.0%, Confianza: 2.5%
12	{default} -> /prod/database/isoc/isoc-tex.html	Soporte: 100.0%, Confianza: 2.5%
13	{default} -> /prod/database/isoc/icyt-firstload.html	Soporte: 100.0%, Confianza: 2.5%
14	{default} -> /prod/database/isoc/isoc-sidebar.html	Soporte: 100.0%, Confianza: 2.5%
15	{default} -> /prod/database/icyt/icyt-tex.html	Soporte: 100.0%, Confianza: 2.5%
16	{default} -> /prod/database/icyt/icyt-firstload.html	Soporte: 100.0%, Confianza: 2.4%
17	{default} -> /prod/database/icyt/icyt-fa.html	Soporte: 100.0%, Confianza: 2.4%
18	{default} -> /prod/database/icyt/icyt-sidebar.html	Soporte: 100.0%, Confianza: 2.4%
19	{default} -> /prod/database/icyt/icyt.html	Soporte: 100.0%, Confianza: 2.4%
20	{default} -> /prod/database/ime/icyt-fa.html	Soporte: 100.0%, Confianza: 1.6%
21	{default} -> /prod/database/ime/ime-tex.html	Soporte: 100.0%, Confianza: 1.6%
22	{default} -> /prod/database/ime/ime-firstload.html	Soporte: 100.0%, Confianza: 1.6%
23	{default} -> /prod/database/ime/ime-sidebar.html	Soporte: 100.0%, Confianza: 1.6%
24	{default} -> /prod/database/ime/ime.html	Soporte: 100.0%, Confianza: 1.6%
25	{default} -> /servicios/revistas_electronicas.html	Soporte: 100.0%, Confianza: 2.1%
26	{default} -> /servicios/docen2.html	Soporte: 100.0%, Confianza: 2.1%
27	{default} -> /cybermetrics/cybermetrics.html	Soporte: 100.0%, Confianza: 1.9%
28	{default} -> /isis/c1.htm	Soporte: 100.0%, Confianza: 1.7%
29	{default} -> /info/info.html	Soporte: 100.0%, Confianza: 1.8%
30	{default} -> /servicios/servicios.html	Soporte: 100.0%, Confianza: 1.5%
31	{default} -> /prod/productos.html	Soporte: 100.0%, Confianza: 1.5%
32	{default} -> /prod/database/isocdc/ime-sidebar.html	Soporte: 100.0%, Confianza: 1.4%
33	{default} -> /prod/database/isocdc/isoc.html	Soporte: 100.0%, Confianza: 1.4%
34	{default} -> /prod/database/isocdc/ime-firstload.html	Soporte: 100.0%, Confianza: 1.4%
35	{default} -> /prod/database/isocdc/icyt-fa.html	Soporte: 100.0%, Confianza: 1.4%
36	{default} -> /prod/database/isocdc/alat-tex.html	Soporte: 100.0%, Confianza: 1.4%
37	{default} -> /redc/redc.html	Soporte: 100.0%, Confianza: 1.4%
38	{default} -> /info/qcindoc.html	Soporte: 100.0%, Confianza: 1.1%
39	{default} -> /isis/01-1.htm	Soporte: 100.0%, Confianza: 1.1%
40	{default} -> /isis/enlaces.htm	Soporte: 100.0%, Confianza: 1.3%
41	{default} -> /investigacion/investigacion.html	Soporte: 100.0%, Confianza: 1.2%
42	{default} -> /orecursos/recinfdoc.html	Soporte: 100.0%, Confianza: 1.2%

Figura 2. Informe de reglas de asociación del software WebMining log sessionizator 7.0

casos, gran facilidad de instalación, opción de hacer públicas las estadísticas), pero también importantes limitaciones y desventajas:

– El grado de personalización es muy limitado, cuando no inexistente. Se ofrecen las estadísticas "as is", sin apenas información de cómo se obtienen ni posibilidad de adecuarlas a necesidades concretas.

Visitantes	Distribución por hora, día, mes y año
	Únicos, repetidos (mismo IP)
Visitas	Hits (total ficheros visitados: html e incorporados), ficheros individuales, páginas visitadas, páginas por visita, directorios visitados
	Ficheros bajados o volcados
Origen	Según IP, ciudad, dominio, o país. Idiomas. Mapas
Intermediarios	Motores de búsqueda, términos utilizados, frases utilizadas
Comportamiento	Ruta de la visita, tiempo utilizado, páginas de entrada y salida
Tecnología	Navegador usado, resolución de pantalla, sistema operativo, velocidad de conexión

Tabla 1. Algunas variables ofrecidas por contadores

– Algunas de las empresas que ofrecen estos servicios no lo tienen como prioritario, y a veces la calidad del mismo deja mucho que desear: cortes de suministro, funcionamiento irregular, pérdida de datos. En casos extremos el servicio puede desaparecer sin previo aviso.

"*StatCounter* y *Google Analytics* son herramientas muy útiles tanto por su facilidad de instalación y coste cero, como por las amplias estadísticas que ofrecen"

– Al contrario que los programas de análisis de ficheros log que apenas tienen problemas para el seguimiento de volcados de ficheros, estos programas no parecen pensados para obtener estadísticas fiables al respecto. En principio habría que añadir el código de control a todos y cada uno de los ficheros "bajables", que generalmente no están en formatos fácilmente editables (*Acrobat*, *Postscript*, *Word*, *Powerpoint*) o que no admiten fácilmente texto html. Además de esta di-

Motigo Webstats webstats.motigo.com	Opentracker www.Opentracker.net
123 count www.123count.com	Realtracker www.realtracker.com
3D Stats www.3dstats.com	Servustats www.servustats.com
AceStats www.acestats.com	Shinystat www.shinystat.com
Addfree Stats www.addfreestats.com	Site Meter www.sitemeter.com
ClustrMaps www.clustrmaps.com	Sitetracker www.sitetracker.com
CounterCentral www.countercentral.com	SiteTrafficStats www.sitetrafficstats.com
Count My Page www.countmypage.com	Statcounter www.statcounter.com
Cq Counter www.cqcounter.com	Stats For Your Site www.statsforyoursite.com
Dataplain Web Stats www.dataplain.com	SuperStats www.superstats.com
Digits Web Counter www.digits.com	Traffic Examiner www.trafficexaminer.com
Easy Counter www.easycounter.com	Traffic File www.trafficfile.com
FreeStats www.freestats.com	Vioclicks www.vioclicks.com
GoldStats www.goldstats.com	W3Counter www.w3counter.com
GoStats gostats.com	WebStat www.webstat.com
Histats www.histats.com	Web-Stat www.web-stat.com
Hitmatic www.hitmatic.com	Webtistic www.webtistic.com
Hitslink www.hitslink.com	WunderCounter www.wundercounter.com
iWebTrack www.iwebtrack.com	Estadisticas Gratis www.estadisticasgratis.com
Mega Stats www.megastats.com	Weboscope www.weboscope.com
NedStat www.nedstat.com	Controlia www.controlia.com
Nextstat www.nextstat.com	Contador de visitas www.contadordevisitas.org
One Stat www.onestat.com	Webcontroler www.webcontroler.es

Tabla 2. Sistemas de contadores y estadísticas gratuitos o semi-gratuitos

ficultad, extrema si consideramos grandes repositorios, los programas no están preparados para tratamientos estadísticos diferenciales de estos datos: no segregan ficheros volcables del resto de páginas, no hacen agrupaciones ni seguimientos específicos, no tienen protocolos para diferenciar visitas de volcados.

Con el fin de mostrar las principales características de estos sistemas se ha realizado un pequeño experimento con dos de las opciones más populares: *Statcounter* y *Google Analytics* utilizados en el seguimiento de visitas del *Ranking Web de Universidades del Mundo*:

http://www.webometrics.info

Statcounter es un programa con una opción gratuita básica que hace estadística de visitas y visitantes ilimitada pero sólo proporciona detalles sobre las 500 últimas visitas.

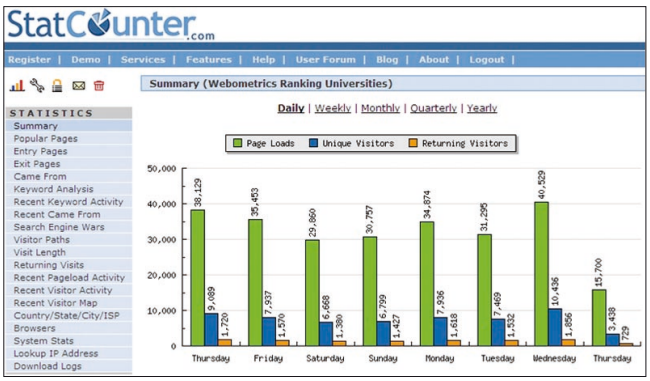


Figura 3. StatCounter (composición)

Google Analytics está basado en el motor *Urchin* y es completamente gratuito, aunque al contrario que el anterior no ofrece una versión pública. Sólo el webmaster tiene acceso a su sofisticada y completa interfaz.

Figura 4. Google Analytics (composición)

El experimento consistió en comparar los resultados obtenidos en dos fechas diferentes: visitas mensuales entre noviembre de 2007 y septiembre de 2008, y visitas diarias entre el 1 y el 16 de octubre de 2008 (siempre ambos inclusive).

Se prestó atención al número de páginas vistas y al número de visitantes, variables comunes tanto a *Statcounter* (SC) como a *Google Analytics* (GA). La correlación entre las cuatro variables fue altísima, entre 0,96 y 0,99, lo que indica que ambos sistemas reflejaron los mismos patrones de uso.

Sin embargo como se muestra en la figura 5, mientras que el número de páginas visitadas (eje Y izquierdo) es prácticamente idéntico en el conteo de ambos sistemas, hay una diferencia en cuanto a volumen en lo que respecta al número de visitantes (eje Y derecho).

Figura 5. Comparativa entre visitantes y páginas vistas según dos sistemas de estadísticas distintos

El número de visitantes de *Statcounter* es aproximadamente un tercio más numeroso que el proporcionado por *Google Analytics*. Puesto que los patrones son muy similares cabe especular que las diferencias son debidas al diferente mecanismo de identificación de visitantes "repetidos" ligados tanto a la identificación del IP visitante como al seguimiento temporal de la visita. En el caso de *Statcounter* dos visitas consecutivas del mismo usuario pueden ser consideradas distintas si ha pasado un plazo de tiempo inferior al que está definido en *Google Analytics*.

En resumen se trata de herramientas muy útiles tanto por su facilidad de instalación y coste cero, como por las amplias estadísticas que ofrecen. En todo caso cabe cuestionar su uso con fines comparativos salvo si se trata del mismo producto y bajo la mismas condiciones. La marca *Google* ofrece una ventaja adicional, ya que podría convertirse en un estándar de facto lo que aumentaría su presencia en el mercado. Si se deciden a abrir los resultados al público, la disponibilidad de información de un mayor número de sedes enriquecería la comparación con datos de todo el mundo y distintos sectores.

Bibliografía

Gomory, S.; **Hoch, R.**; **Lee, J.**; **Podlaseck, M.**; **Schonberg, E.** Analysis and visualization of metrics for online merchandizing. En: *WebKDD*, Springer, San Diego, CA, 1999.

Thelwall, M. "Web log file analysis: backlinks and queries". *Aslib proceedings*, 2001, n. 53, pp. 217-223.

Peters, T. A. "The history and development of transaction log analysis". *Library hi tech,* 1993, n. 42, pp. 41–66.

Kurth, M. "The limits and limitations of transaction log analysis". *Library Hi Tech*, 1993, n. 42, pp. 98–104.

Spiliopoulou, M. "Web usage mining for web site evaluation". *Communications of the ACM*, 2000, v. 43, n. 8.

Klevecz, B. "The whole EST catalog". *Scientist*, 1999, v. 12, n. 2, p. 22.

Cooley, R.; **Mobasher, B.**; **Srivastava, J.** "Data preparation for mining world wide web browsing pattern". *Knowledge and information systems*, 1999, v. 1, n. 1, pp. 5-32.
http://maya.cs.depaul.edu/~mobasher/papers/webminer-kais.pdf

Cooley, R.; **Mobasher, B.**; **Srivastava, J.** "Web mining: information and pattern discovery on the world wide web". En: *Proceedings of the 9th IEEE International*, 1997.

Pitkow, J. "In search of reliable usage data on the WWW". En: *Sixth international world wide web conference*, 1997, Santa Clara, CA, pp. 451-463.

Catledge, L.; **Pitkow, J.** "Characterizing browsing behaviors on the World Wide Web". *Computer networks and ISDN systems*, 1995, v. 27, n. 6, pp. 1065-1073.

***José-Luis Ortega**, División de Programación Científica, Vicepresidencia de Ciencia y Tecnología, CSIC, Serrano 113.*
28006 Madrid, España.
jortega@orgc.csic.es

***Isidro F. Aguillo**, Laboratorio de Cibermetría, CCHS, CSIC, Albasanz 26-28.*
28037 Madrid, España.
isidro.aguillo@cchs.csic.es