

# Minería de uso de la web mediante huellas y sesiones

Julio Villena<sup>1</sup>, Emma Barceló<sup>2</sup>, Juan Ramón Velasco<sup>2</sup>

<sup>1</sup>DAEDALUS, S.A. – Paseo de las Delicias 31, 3º  
28045 Madrid (España)  
jvillena@daedalus.es

<sup>2</sup>Depto. Ingeniería de Sistemas Telemáticos – ETSIT – Universidad Politécnica de Madrid  
Av. Ciudad Universitaria, s/n – 28045 Madrid (España)  
{ebarcelo, juanra}@gsi.dit.upm.es

**Resumen.** Dentro de las aplicaciones de la minería de datos en la web, se encuentra un tipo concreto denominado minería de uso de la web, cuyo objetivo es extraer información y conocimiento útil sobre el tráfico en un determinado servidor. En estos estudios se emplea normalmente como métrica el número de accesos a los objetos contenidos, que quedan almacenados en el log del servidor. Estos registros están distorsionados por la existencia de cachés en la red: aunque un cliente visite una página determinada, la petición puede no llegar al servidor por estar almacenada en un servidor intermedio, y por tanto la petición no registrarse en el log. Aquí se presenta el método de las huellas para realizar las mediciones, que soluciona el efecto de las cachés. Por otro lado, los accesos sólo miden la carga del servidor, no el número de visitantes. Esta medida es la sesión de usuario, que se plantea como una métrica más apropiada para medir la audiencia en la red.

## 1 Introducción

La minería de uso de la web, cada día más popular y extendida, pretende descubrir correlaciones y tendencias significativas en todo tipo de información relacionada con la web aplicando las técnicas y algoritmos de la minería de datos. El análisis del tráfico de acceso a un determinado servidor web, previamente registrado de una manera apropiada, puede ayudar, por una parte, a entender el comportamiento y hábitos de los clientes/usuarios del servidor y, por otra, a diseñar adecuadamente la estructura de la web o mejorar el diseño de esta inmensa colección de recursos.

En la minería de uso de la web, según las aplicaciones de descubrimiento, hay dos aplicaciones principales. La primera de ellas es la *búsqueda de patrones de acceso general* [8] [9] [10] que analiza el tráfico para entender los patrones de acceso y comportamiento habitual de los clientes/usuarios y sus tendencias, para reestructurar el sitio ubicando los contenidos de forma más accesible o para dirigir a los usuarios de la web hacia lugares concretos de la misma. La segunda tendencia es la *búsqueda de uso personalizado* [8] [11], que analiza las tendencias individuales de cada visitante de la web para personalizar o adaptar dinámicamente la información del sitio web, la pro-

fundidad de su estructura y/o el formato de sus recursos a cada visitante en cada momento, según el patrón de acceso que exhiba.

Como en cualquier otra aplicación de minería de datos, el éxito del proceso depende del conocimiento que se descubre y su disponibilidad en un momento determinado, su accesibilidad por los usuarios que realmente lo van a emplear y, sobre todo, su validez y fiabilidad. Por ello es muy importante medir de forma exacta y precisa el tráfico sobre un servidor web, que son los datos que se emplean como entrada en cualquier proceso de extracción de conocimiento sobre la web.

Desde los primeros tiempos, los servidores web tienen la capacidad de registrar en un fichero log los accesos que reciben de los clientes y los estudios se han basado tradicionalmente en estadísticas más o menos avanzadas sobre estos accesos. Pero estas mediciones tienen dos problemas fundamentales.

Por un lado, la existencia de cachés en la red hace que aunque un cliente visite una página determinada, la petición pueda no llegar al servidor por estar almacenada en un servidor intermedio, y por tanto la petición no queda registrada en el log. Esto produce que el tráfico y uso medido sea inferior al real, distorsionando la validez de los estudios y de los controles de audiencia en la web. Aquí proponemos otro método de registro del tráfico, el método de las huellas, para evitar este problema.

Por otra parte, las métricas (unidades) que se emplean tradicionalmente en los análisis son los accesos. De esta manera en realidad se está midiendo la carga del servidor –en el sentido de cantidad de objetos accedidos– y no la audiencia –en el sentido del número de personas que visitan el sitio–, puesto que un mismo usuario podría acceder varias veces a la misma página cuando está navegando por el sitio web, o incluso recargando una de ellas por alguna razón. En este documento se propone otro tipo de métrica: el concepto de sesión, como agrupación de todas las páginas que visita un mismo usuario.

En el apartado 2 se exponen los fundamentos del método de las huellas y en el apartado 3 se presenta la definición y la utilidad de adoptar la sesión de usuario como métrica del uso, calculada a partir de los accesos. El apartado 4 recoge una evaluación realizada sobre un sitio web real empleando un sistema de minería de uso de la web en tiempo real, Lawerinto Miner® de DAEDALUS [1] [2], basado en huellas y sesiones, que demuestra la conveniencia de estos conceptos. Por último, en el apartado 5 recogemos las conclusiones y los trabajos que quedan pendientes como vías futuras.

## 2 Registro del tráfico

La unidad que tradicionalmente se emplea como medida del tráfico de un sitio web es el **acceso**, **hit** o **impresión**. Un acceso es una petición individual de carga de un objeto (páginas, documentos, imágenes, vídeos...) que recibe el servidor por parte del cliente. Así, el tráfico de un determinado sitio web es la carga del servidor, es decir, el número de objetos servidos en un determinado período de tiempo (accesos/tiempo).

## 2.1 Log del servidor

Desde los primeros tiempos de la red, los servidores web registran de forma incremental todos los accesos a objetos recibidos desde los clientes en el denominado **log del servidor**. Este log del servidor es el que se emplea para los análisis del tráfico y uso siendo periódicamente procesado y analizado por el responsable del sitio web para extraer información sobre el uso de su servidor.

El log se almacena en ficheros de texto plano (ASCII) en diferentes formatos (NCSA, W3C y Microsoft IIS), aunque algunos servidores pueden almacenarlo directamente en una base de datos. La información que se almacena depende también de cada servidor y su configuración particular, pero en general los campos incluidos habituales son: la fecha y hora de la petición, la máquina (y/o dirección IP) del cliente, el método de acceso (GET, PUT...), el fichero accedido (URL), el resultado de la petición, el tamaño en bytes de los datos recibidos y devueltos, la identificación del cliente (*User-Agent*), cookies y la página de procedencia (*Referrer*).

Un serio problema al usar el log del servidor son los accesos que no son registrados, ya que mecanismos como las cachés (locales o compartidas) y los servidores proxy [3] [7] pueden distorsionar gravemente la imagen global de los movimientos del usuario por el sitio web. Un objeto listado sólo una vez en un histórico de accesos puede haber sido accedido muchas veces por distintos usuarios, sin que las peticiones lleguen al servidor al estar el objeto almacenado en una de estas cachés intermedias.

Las páginas dinámicas no se ven afectadas porque (en principio) no son almacenadas en las cachés. Para el caso de objetos estáticos, HTTP incluye directivas de control de caché [4] que fuerzan que los objetos (en teoría) no se almacenen en las cachés (pragma=no-cache en HTTP/1.0, cache-control=public/private/no-cache/no-store en HTTP/1.1), cabeceras *Expires* y *Last-modified*, etc. Otros métodos [8] son la limpieza de caché, registro explícito de usuario, etc., pero ninguno se encuentra libre de problemas. Es más, de los estudios realizados se desprende que hay algunos casos en que según el tipo de enlace con el que un navegador accede a un objeto, este objeto es siempre almacenado en la caché del navegador. En particular sucede cuando se trata de una imagen, ya sea por diseño del navegador o por *bugs*.

El método más habilidoso para evitar que un objeto se almacene en una caché es que su URL sea irrepetible e único cada vez de que se acceda. Para ello a cada URL se le añade como parámetro un número aleatorio (por ejemplo, "url.htm?n=NNNN") que se ignora. Esto solo sería posible si la página se genera dinámicamente (en el servidor o en el cliente, por ejemplo con javascript).

A lo largo de este trabajo se han desarrollado métodos para intentar la reconstrucción del log e inferir las referencias que faltan, incluyendo, además de la información temporal, el uso de información sobre la topología del sitio y ficheros históricos de envíos, pero los resultados han sido escasos y poco concluyentes.

Otro inconveniente no desdeñable para emplear el log del servidor es su disponibilidad. Actualmente muchas organizaciones tienen albergados sus sitios web en servidores externos, compartidos entre muchos, y no pueden disponer de los ficheros log.

## 2.2 Huellas

La solución alternativa que proponemos al registro en un log de servidor es el empleo de **huellas**. Una huella es un rastro o marca que queda registrada por parte del usuario al acceder a un determinado objeto. Las huellas se basan en la inclusión dentro de las páginas que se quieren controlar de una referencia (enlace) a un elemento adicional, que va a provocar una nueva petición por parte del cliente para acceder a ese elemento, con el efecto colateral de registrar el acceso a la página que la contiene.

Este elemento adicional (huella) es un enlace a una página dinámica (en forma de CGI, ASP, etc.) que recibirá (como parámetros del URL) aquella información que se quiera registrar sobre la página que se está visitando y la almacenará, preferiblemente en una base de datos. Esta página puede estar (y habitualmente lo estará) en un servidor distinto al que se quiere controlar. En este punto se introduce el concepto de **webmart** (o webhouse) como almacén de la información de uso de un sitio web [6].

Dado que los objetos que pueden incluir referencias a otros objetos son únicamente las páginas HTML, este método sería válido en principio para controlar páginas HTML, aunque también se podría controlar el resto de objetos (documentos, vídeos...) si se accediera a ellos mediante una página dinámica que primero llamara internamente a la huella y que luego devolviera el objeto solicitado.

El proceso es el siguiente: un usuario hace una petición de una página a un servidor y éste se la entrega. Cuando el navegador presenta la página en pantalla, accede al enlace de la huella, enviando información del visitante que se incorpora al webmart, en la tabla operacional. Esta tabla contiene datos interesantes como el objeto (URL) accedido, la fecha y hora, la máquina (o su dirección IP) origen, el método de acceso, la página de procedencia, las cookies enviadas, la identificación del navegador cliente, la resolución de pantalla del cliente, etc.

La primera ventaja es que ya no es necesario el log del sitio a controlar, porque los accesos se registran en otro servidor. Además, el análisis puede hacerse en tiempo real, e incluso el servicio puede ser prestado por una organización independiente.

## 2.3 Tipos de huella

La forma de incluir la huella y la información que puede almacenar depende de la tecnología disponible y de los tipos de usuarios cuyos accesos se pretendan controlar. En general en las visitas a un sitio web hay que distinguir entre el tráfico/uso generado por personas (objetivo de estudios de comportamiento, campañas de marketing...) y el tráfico/uso artificial generado por robots cuya finalidad normal es recorrer las páginas pero sin presentarlas a ningún usuario.

En el primer caso, los usuarios emplean programas navegadores para acceder a los objetos del servidor, cuya funcionalidad y comportamiento son generalmente bien conocidos. En el caso del tráfico/uso artificial [5], la forma de acceso y el comportamiento depende de cada robot en particular, existiendo miles de ellos en la actualidad.

Hay robots “amistosos” que siguen el estándar de buen comportamiento de robots, identificándose ante el servidor (con la cabecera HTTP “User-Agent”), accediendo al fichero robots.txt de cada servidor que visitan y siguiendo sus instrucciones. Pero

otros robots son “maliciosos” y sus accesos tienen un comportamiento anómalo en el sitio web, con finalidad diversa. Por ejemplo, no se identifican, realizan peticiones de páginas no autorizadas, peticiones de páginas inexistentes, aquellas que realizan operaciones no permitidas sobre las páginas del servidor. Es importante distinguirlos de los robots “amistosos”. La detección puede inducirse a partir de las sesiones de robots “amistosos”, por ejemplo mediante una clasificación borrosa (fuzzy) a posteriori con el comportamiento detectado.

Como se dijo anteriormente, la página de la huella es una página dinámica, que no se ve, en principio, afectada por las cachés. En cualquier caso debe incluir todos los mecanismos definidos para control de caché, y lo ideal si se puede es emplear el método del URL irrepensible. Se ha mencionado también que el enlace a la huella debe tener como parámetros información sobre la página que se quiere controlar.

Hay información que el navegador envía en la petición a la página de la huella que no sería necesario incluir en los parámetros: el URL de la página (es la página de procedencia -referrer- de la huella), la dirección IP de acceso (la misma que la de acceso a la página), la fecha y hora de la petición (que coinciden aproximadamente con las del acceso a la página que tiene la huella, despreciando su tiempo de carga de análisis de enlaces), tecnología del cliente (navegador o robot), cookies, autenticación de usuario del cliente, etc. Sin embargo, por ejemplo, no hay forma de obtener la página de procedencia de la página que contiene la huella, que debe pasarse como parámetro.

El URL es lógicamente el parámetro más importante. En algunos casos cuando el cliente no proporciona al servidor donde reside la huella la información de la página de procedencia (que coincidía con el URL), como casi todos los robots hacen, es necesario pasar explícitamente el URL como parámetro. Además hay que mencionar que podría ser interesante no emplear el URL real a favor de un alias (estático) al mismo (más representativo, por ejemplo, cuando el URL tiene muchos parámetros y es muy largo) o para agrupar lógicamente una serie de URLs en un único nombre.

La huella óptima es aquella para detectar **navegadores** que soportan **javascript**. En este caso, los navegadores envían información de la página de procedencia al servidor (no hace falta el parámetro URL), se puede emplear el método del URL irrepensible, y se usa un enlace de tipo <IMG SRC=...> (imagen) porque el navegador va a descargar automáticamente todas las imágenes incluidas al presentar la página. La huella devolverá una imagen transparente de 1x1 píxeles invisible en la pantalla. Esta huella es la denominada **huella dinámica de cliente** (Figura 1).

```
<SCRIPT type="text/javascript">
  document.write("<img src='http://servidor/Huella.asp?ref=" +
    escape(document.referrer)+"&rnd=" + (new Date()).getTime() + "'>");
</SCRIPT>
```

**Fig. 1.** Huella para navegadores con soporte javascript

En la huella para detectar navegadores que no soportan javascript se podrá omitir el URL, pero no se podrá obtener la información de la página de procedencia de la página visitada. Esta huella es una **huella estática** (Figura 2).

```
<NOSCRIPT>
  
</NOSCRIPT>
```

**Fig. 2.** Huella para navegadores sin soporte javascript

Para detectar sesiones de robots (suponiendo que no envían información de la página de procedencia) hay que tener en cuenta que en general los robots sólo están interesados en páginas HTML y quizás en ciertos tipos de documentos, pero habitualmente no recuperan las imágenes. Por ello se sustituye el enlace de la imagen por un <A HREF=...>. Además un robot no suele soportar javascript, así que hay que incluir el parámetro URL. Por tanto la huella para robots es también estática (Figura 3).

```
<NOSCRIPT>
  <a href="http://servidor/Huella.asp?url=url.htm"></a>
</NOSCRIPT>
```

**Fig. 3.** Huella para robots

La inclusión del parámetro URL puede ser una grave molestia en el caso de que el sitio web tenga muchas páginas puesto que hay que poner algo distinto en cada una de ellas. De todas maneras, ya que puede suponerse que un robot en general va a recorrer todo el sitio web y visitando una única vez cada página, para el estudio puede bastar simplemente con detectar la entrada de un robot, poniendo la huella únicamente en la página de inicio (que casi seguro será visitada). Para el estudio de la inducción de sesiones maliciosas se hace necesario poner la huella estática en todas las páginas, para posteriormente poder inferir comportamientos semejantes entre sesiones no identificadas como robots previamente, y sesiones de robots amistosos.

Debe considerarse que si el comportamiento del robot es limitado, es decir, sólo recorre los enlaces del sitio web y no pide enlaces externos, entonces la huella que venimos comentando no es válida si no está en el mismo servidor que las páginas HTML, y la solución pasa por implementar técnicas de redirección de llamadas. Es decir, la inclusión dentro de cada página en la que se quiera controlar el acceso de los robots de un código que llame internamente al programa que registra el acceso en el webmart. Este problema está actualmente en estudio.

Por último, está la llamada **huella dinámica de servidor**, que es la incluida en una página dinámica cuando se accede a ella y generada dinámicamente por ésta. Para el cliente es una huella estática, pero con todos los parámetros incorporados (Figura 4).

```
<a href="http://servidor/Huella.asp?url=$url&ref=$referrer&rnd=NNN"></a>

```

**Fig. 4.** Huellas dinámicas de servidor

De todo esto se deduce que según la tecnología de acceso al objeto que se emplee (si envía o no información al servidor, si soporta javascript o no) la funcionalidad que se puede obtener con la huella es más o menos rica.

El código de la(s) huella(s) se inserta en todas y cada una de las páginas que se desee controlar, entre las etiquetas <BODY> y </BODY>, y preferiblemente al principio de la página por si el usuario interrumpe la descarga. Lo más cómodo es insertarla en un pie de página común a todas las páginas web del sitio (huellas dinámicas).

## 2.4 Combinación log del servidor + huellas

La tabla 1 presenta una comparativa de las ventajas e inconvenientes de cada una de las opciones. Como casi siempre, se deduce que la información más interesante se obtiene con la combinación de ambas, esto es, la información obtenida con las huellas con la información registrada en el log, eliminando posibles efectos de caché y completando la información de uno y otro método. Por ejemplo, con sólo huellas dinámicas no se podrían detectar sesiones de robot, pero con el log sí. Y al revés, con el log no se pueden diferenciar de forma directa qué accesos son de personas y cuáles son de robots, pero con las huellas sí.

**Tabla 1.** Comparativa entre registros en el log del servidor y con huellas.

	Log	Huella estática	Huella dinámica cliente	Huella dinámica servidor
<b>Detección de robots</b>	Sí (amistosos)	Sí	No	Sí
<b>Sobrecarga o retardo en la carga de la página</b>	No	Sí (muy muy poca)	Sí (muy poca)	Sí (muy muy poca)
<b>Modificación del diseño web</b>	No	Sí (muy pequeño)	Sí (muy pequeño)	Sí (pequeño)
<b>Contempla el efecto de la caché</b>	No	No	Sí	Sí
<b>Tipo de objetos</b>	Todos		HTML (el resto con modificaciones)	
<b>Tipo de información</b>	La de las peticiones HTTP		La de las peticiones HTTP más la del navegador web del cliente	
<b>Procesamiento</b>	Diferido, batch		Tiempo real	
<b>Alojamiento</b>	En el servidor		En webmart externos o internos	

## 3. Empleo de la sesión como métrica del uso

El principal problema de emplear los accesos como métrica del uso es que su valor se incrementa artificialmente por el efecto de que un mismo cliente acceda al mismo objeto varias veces, ya sea en el transcurso normal de su navegación por el sitio web o por el refresco de la página. Esto significa que el uso así medido queda afectado por la estructura de los enlaces entre páginas del sitio o por problemas de descarga de páginas u otros motivos que causan peticiones de refresco de objetos, lo que es inadecuado.

### 3.1 Sesiones

Una métrica más adecuada es aquella basada en el número de visitantes individuales de un sitio web, independientemente del número de páginas que visite, que es el concepto de sesión. Una **sesión** [6] de usuario está formada por el conjunto de objetos consultados por un mismo visitante durante una misma visita al sitio web. De esta manera, el uso de un sitio se mide con el número de sesiones de usuario en un período de tiempo (usuarios/tiempo) y no con la carga del servidor (accesos/tiempo).

Una sesión de usuario está formada por el conjunto de páginas consultadas por un visitante durante una sola visita al sitio web. Habitualmente los criterios de corte de sesiones se establecen por tiempo, siendo frecuente un umbral de inactividad de 10 minutos. En este estudio se aporta, además del criterio de inactividad, el hecho de que superado este umbral, la nueva página visitada por el presuntamente mismo visitante sea accedida desde la última página registrada. Es decir, se rompe la sesión si el visitante no viene de la página anteriormente vista, sino que ha accedido desde otra no registrada.

De nuevo, la combinación de ambas métricas aporta un mejor conocimiento de la actividad del sitio web que cada una por separado, puesto que número de accesos refleja el número total de impresiones de la página, por tanto, la carga del servidor, mientras que el número de sesiones de usuario indica los visitantes únicos que han accedido, por tanto, el efecto (impacto en usuarios) obtenido con el servidor.

## 4 Medidas en un sistema real

Para la comprobación de esta propuesta se ha empleado un sistema comercial de minería de uso de la web en tiempo real llamado Lawerinto Miner® [2] para medir el uso de otro sitio web. Este sistema, descrito en [1], se basa en los conceptos de huella y sesiones aquí presentados como métrica del uso.

El efecto de las cachés se puede estudiar a través del uso comparativo de las huellas dinámicas y estáticas. Para ello hemos tomado un sitio web universitario real (<http://www.gsi.dit.upm.es/~juanra>) en el que se han insertado dos huellas en cada página, una dinámica y otra estática, y se ha registrado el uso durante un periodo de 16 días (21/6/2001 al 6/7/2001). Hay estadísticas que no se pueden ver con la huella estática, como la procedencia de las visitas o las consultas que llevan a los usuarios al sitio web consultado, ya que estos parámetros no pueden enviarse con esta huella.




El efecto de las cachés se muestra en la Tabla II. La pérdida de accesos por el efecto de las cachés es considerablemente superior a la pérdida de sesiones, (27,34% frente a 7,09%), con lo que se corrobora la mejor elección de las sesiones como métrica. Además, se puede observar cómo las dos medidas promedio también disminuyen significativamente: el número de páginas promedio porque las páginas han sido almacenadas en cachés y por tanto las sesiones tendrán menos páginas vistas, y la duración promedio ya que el efecto caché influye directamente en el tiempo de descarga de las páginas, además del efecto de la disminución del número de páginas.






**Tabla 2.** Efecto de las cachés sobre un sitio web.

	Accesos	%Accesos	Sesiones	%Sesiones	Pág pro- medio	Dur pro- medio
<b>Estática</b>	1445	72,68%	800	92,91%	1,81	48,90
<b>Dinámica</b>	1988	100%	861	100%	2,31	96,00
<b>Pérdidas</b>	543	27,32%	61	7,09%		

Otro efecto interesante es la modificación de las sesiones. Debido al efecto de las cachés, las sesiones se rompen, ya que si una página intermedia ha sido almacenada en la caché, la siguiente página que se pide al servidor no se registra en el log, quedando truncada la sesión. Este efecto conlleva la modificación del número de sesiones que tienen como página de entrada una determinada URL, que pueden incluso llegar a ser mayores con la huella estática que con la dinámica (Figuras 5 y 6).

PRIMERA PÁGINA VISITADA EN LA SESIÓN		
Página	Número de sesiones	
/~juanra/scouts/scoutsgu.html	163	
/~juanra/scouts/basecamp.html	124	
/~juanra/isof/0001/traspas.html	103	

**Fig. 5.** Páginas de entrada a la sesión con la huella estática

PRIMERA PÁGINA VISITADA EN LA SESIÓN		
Página	Número de sesiones	
http://www.gsi.dit.upm.es/~juanra/scouts/scoutsgu.html	161	
http://www.gsi.dit.upm.es/~juanra/scouts/basecamp.html	112	
http://www.gsi.dit.upm.es/~juanra/isof/0001/traspas.html	107	

**Fig. 6.** Páginas de entrada a la sesión con la huella dinámica

En la tabla siguiente se muestra un ejemplo de detección de sesiones de robots (con más del 50% de los accesos a www.gsi.dit.upm.es entre el 1/6/2001 y el 15/6/2001).

**Tabla 3.** Accesos de Robots y accesos totales del 1/6/2001 al 15/6/2001

	Accesos	%Accesos
<b>Robots (que se identifican)</b>	43324	56,74%
<b>No robots</b>	33042	43,26%
<b>Total</b>	76366	100%

## 5 Conclusiones y trabajos futuros

Se ha corroborado que con el sistema de medición del tráfico y uso de la web de huellas presentado aquí se consigue evitar el efecto de las cachés, tanto del navegador cliente como de la jerarquía de servidores de la red. Además, la métrica encontrada de

sesión de usuario que no conlleva un corte estricto de la visita de un cliente sino que tiene en cuenta la consecución lógica de la visita, sea cual sea el tiempo de inactividad, proporciona una visión más acertada del uso de la web. La detección de sesiones de robots con comportamiento limitado, es decir, que no recorren enlaces externos al sitio web esta siendo estudiada. La identificación de sesiones maliciosas mediante técnicas borrosas se propone como trabajo futuro, a partir de la caracterización de sesiones de robots.

## Referencias

- [1] E. Barceló, J.Villena, J.R.Velasco; Desarrollo de un Sistema de Minería de Uso de la Web en Tiempo Real. CAEPIA 2001.
- [2] DAEDALUS – Data, Decisions and Language, S.A. Lawerinto Miner® <http://www.lawerinto.com>
- [3] B. Duska, D. Marwood, and M. J. Feeley. The measured access characteristics of World Wide Web client proxy caches. USENIX Association. 1997.
- [4] The Cache Now! Campaign; <http://vancouver-webpages.com/CacheNow/>
- [5] D. Eichmann. Ethical Web Agents. Computer Networks and ISDN Systems. Vol. 28, 1 a 2, pag. 127 a 136, 1995.
- [6] R. Kimball and R. Merz. The Data Webhouse Toolkit. John Wiley and Sons, Inc., 2000.
- [7] A. Rousskov, "On Performance of Caching Proxies", NoDak, February 1998.
- [8] R. Cooley , J. Srivastava, B. Mobasher, Web Mining: Information and Pattern Discovery on the World Wide Web, Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97). November 1997.
- [9] J. Borges and M. Levene: *Data mining of User Navigation Paterns*, Proceedings WEBKDD'99. pág. 92-111. 1999.
- [10] Ramakrishana Srikant; Research Report: Mining Sequential Patters: Generalizations and performance improvements. EDBT. Avignon, France, 1998.
- [11] J. C. Bezdek y S. K. Pal (eds.). Fuzzy Models for Pattern Recognition: Methods That Search for Structures in Data, IEEE Press, Piscataway, NJ, USA 1992.