# Minería de Datos para análisis del uso de sitios web

Juan Ramón Velasco Pérez y Luis Magdalena Layos

Departamento de Automática, Universidad de Alcalá de Henares
Escuela Politécnica, Campus Universitario N-II, km. 31,5, 28871 Alcalá de Henares, Madrid
juanra@aut.uah.es
Departamento de Matemática Aplicada a las TT.II., Universidad Politécnica de Madrid
ETSI Telecomunicación, Ciudad Universitaria s/n, 28040, Madrid
llayos@mat.upm.es

**Resumen**: Las técnicas de minería de datos son aplicables en cualquier entorno para el que dispongamos de un volumen de datos elevado. Estas técnicas nos permiten descubrir conocimiento oculto no trivial, estableciendo relaciones entre los atributos que describen los datos utilizados. En los sitios web se acumula una gran cantidad de datos, ya que quedan registradas todas las peticiones de objetos, páginas, gráficos, etc., que el sistema ofrece. Además del análisis estadístico, técnica tradicional para analizar los datos de un sitio web, la minería de datos puede ofrecer técnicas aceptables para conocer mejor a los visitantes que acceden a sus contenidos.

## 1 Introducción

El análisis del comportamiento de los visitantes en un sitio web es una de las actividades más interesantes para la aplicación de técnicas de minería de datos en el mundo real. Por un lado, se cumple una de las condiciones que deberíamos considerar como fundamental para pensar en la aplicación de este tipo de técnicas: existen conjuntos enormes de datos sobre los que trabajar. Por otro, esos datos no son siempre fiables, por lo que el reto de obtener buenos datos de partida no tiene solución trivial, como veremos más adelante. En tercer lugar, es posible obtener algunos resultados (algo de información) aplicando técnicas diferentes a la minería de datos, más concretamente, estadísticas. Sin embargo, el uso de técnicas propias del aprendizaje automático complementa esa información, permitiéndonos descubrir conocimiento, oculto hasta ese momento.

Dentro del ámbito de la minería de datos existe un estándar de cierta implantación (CRISP-DM [1, 2]). En el presente artículo se considerarán las cuatro primeras fases en que el estándar divide un proyecto de minería de datos (las dos últimas, evaluación y aplicación son propias de cada proyecto real, y no las trataremos). En todo caso, no vamos a entrar en la descripción detallada de cada una de ellas, siguiendo todos sus pasos, sino que nos van a servir para dar un orden lógico al desarrollo del trabajo. En primer lugar, analizaremos el ámbito de aplicación: donde y en qué entorno tiene sentido este trabajo. En segundo lugar, veremos qué datos tenemos disponibles, y describiremos el tratamiento que debemos hacer para poder descubrir nuevo

conocimiento. A continuación, describiremos las posibles técnicas de minería de datos que son relevantes para la aplicación que nos ocupa. Por último, trataremos de extraer algunas conclusiones y definir líneas de trabajos en las que profundizar en el futuro.

## 2 El análisis de sitios web

Hoy en día, todas las organizaciones gastan mucho tiempo, esfuerzo y dinero en la creación de un sitio web; sin embargo muy pocas de ellas prevén realmente el proceso posterior de gestión, mantenimiento, mejora y explotación del mismo. La Web es completamente diferente a otros medios de difusión de información por dos cuestiones fundamentales:

- La Web es anónima. Normalmente no conocemos quién accede a nuestro sitio web, y en muchos casos, ni siquiera sabemos si dos accesos diferentes han sido realizados por una misma persona o no.
- La Web es interactiva. La información presentada al visitante del sitio web no es lineal, sino que trasladamos al propio usuario la responsabilidad de adquirir la información a su gusto. Esto supone que cada visitante verá el sitio web de una forma diferente, que no tiene por qué coincidir con el diseño que del mismo se haya realizado.

Sin duda, conocer cómo se comportan los visitantes de un sitio web es una de las actividades imprescindibles para el responsable de información de la entidad. Pero hemos de tener en cuenta que, debido a los dos factores anteriores, es necesario buscar patrones de comportamiento generales, más que tratar de analizar el comportamiento de cada visitante en particular.

Cuando nos sentamos frente a un programa (de los muchos que hay) que nos ofrece un análisis del uso que se hace de un sitio web, lo que solemos encontrar es un sistema capaz de generar un informe estadístico de los accesos. Esta es la situación más extendida y el punto de partida de las expectativas de los receptores de los análisis que generemos: normalmente lo que espera obtener el responsable del sitio web es un conjunto de estadísticas sobre los días y las horas a las que más visitantes acceden, las páginas más visitadas, etc.

El objetivo de los autores es hacer uso de técnicas propias de la minería de datos que complementen y mejoren sensiblemente la información obtenida por las técnicas estadísticas.

## 3 Los datos, la materia prima

Cada vez que un visitante visualiza una página, el servidor web se encarga de entregarle los contenidos que la forman, de manera que el navegador los componga y puedan ser vistos correctamente. Por cada una de esas peticiones, el servidor escribe una línea en un fichero de log, en la que deja constancia de la fecha y la hora de la petición, la dirección IP de la máquina que ha hecho la solicitud, el elemento

solicitado, un código de estado (si ha habido algún error o si el objeto fue entregado correctamente), e incluso la página desde la que se ha realizado la solicitud, en caso de que ésta se hiciera "pinchando" en un enlace de un navegador. Ya que una página web se compone no solo del texto de la página, sino también de imágenes, sonidos y otros objetos, el fichero de log registrará una línea por cada uno de los objetos que se encuentren en nuestro servidor, por lo que la solicitud de una página implica varias líneas en el registro.

Este fichero de log es la fuente de datos más extendida. Sin embargo, estos ficheros presentan un problema que es realmente relevante: 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.

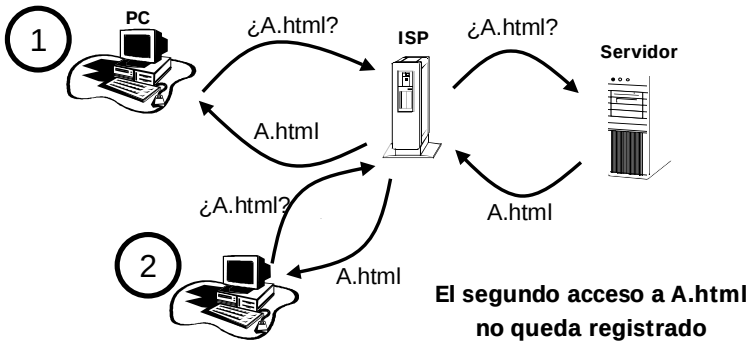



**Fig. 1.** Efecto de las cachés de los proveedores de servicios de Internet

La figura 1 muestra el efecto de las cachés de los proveedores de Internet (ISP): cuando un visitante accede a nuestro sitio web y solicita una página concreta (1), lo hace a través de su proveedor de Internet. Este proveedor se encarga de realizar la petición a nuestro servidor, que le devuelve la página para que se la haga llegar al solicitante. El problema viene cuando un nuevo visitante (2) nos solicita esa misma página a través del mismo proveedor. El efecto es que el ISP guardó una copia de la página en su caché, y sin realizar una nueva petición al servidor, devuelve su copia al visitante. Puesto que no hay petición real, el servidor no puede registrar el acceso del visitante: una visita menos que contabilizamos.

El efecto de esto es que el tráfico y uso medido es inferior al real, distorsionando la validez de los estudios y de los controles de audiencia en la web. Además este efecto se distribuye de forma no uniforme entre los usuarios, las páginas y los servidores falseando aún más los resultados, por lo que una aproximación del tipo “se pierden el 25% de todas las peticiones” no es factible.

Por otro lado, el uso de sistemas automáticos para la recolección o indexación de información es cada día más importante y supone una mayor cantidad de tráfico. Estos sistemas (llamados popularmente “robots de Internet” o simplemente “bots”) realizan peticiones de páginas de forma totalmente análoga a como lo hace un usuario utilizando un navegador, y utilizan la información de los enlaces (links) de esas páginas para realizar nuevas peticiones. En este proceso recopilan, ordenan, buscan o

realizan la función para la que fueron creados. El problema asociado a estos sistemas es que su patrón de navegación no suele reflejar el patrón de un usuario “normal” pero en general en los ficheros log no es posible distinguir unos de otros, con lo que pueden también falsear los resultados obtenidos.

El efecto de todo este falseamiento de datos es que los informes estadísticos no son en absoluto fiables. Pero su influencia en la aplicación de técnicas de minería de datos es aún peor: si el uso de datos erróneos genera resultados erróneos en todo proceso, el aprendizaje a partir de datos erróneos hará que todos nuestras conclusiones sean inexactas. Por este motivo, la obtención de datos fiables se vuelve un elemento imprescindible.

La solución que se asume como más segura para ambos problemas es el empleo de huellas [3]. 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. Estas huellas son capaces de obtener la misma información que se almacena en el fichero de log cada vez que el visitante accede a la página, eliminando directamente el registro de las peticiones de los objetos contenidos en la misma (gráficos, sonidos, etc.) De esta forma sólo tenemos una línea de registro por cada petición de página.

Sea cual sea el mecanismo empleado para la obtención de los datos, al final de esta fase tenemos un fichero repleto de registros, cada uno de los cuales contiene información sobre la petición de una página o un objeto contenido en la misma. Aunque la información registrada depende de la configuración del servidor web, en el caso de querer obtener datos sobre los que realizar un proceso de minería de datos, es conveniente registrar el mayor número posible de características de los accesos: dirección IP desde la que se accede, fecha y hora, identificador del objeto (página) accedido, código de error, tamaño del fichero enviado, página desde la que se ha seleccionado este objeto, etc.

## 4 Procesado de los datos

La tercera fase del estándar CRISP-DM es el procesado de los datos: una actividad encaminada a preparar los datos para que puedan ser utilizados por los diferentes algoritmos de minería de datos. Si los dejamos tal y como son recogidos, difícilmente podremos hacer algo más que calcular estadísticas sobre las páginas más visitadas, las distribuciones de accesos por días y horas, etc. Para poder ir más allá, es necesario trabajar previamente con los datos y generar información adicional.

En primer lugar, es interesante conocer qué páginas han sido visitadas por una misma persona. Dicho de otro modo, cuando un visitante accede al sitio web, solicita un conjunto de páginas. Agrupar esas páginas en sesiones [4] nos permitirá conocer mejor cómo son los visitantes del sitio web.

Pero, ¿Cómo agrupamos las peticiones en sesiones? A un sitio web llegan peticiones simultáneas de muchos visitantes, por lo que el fichero de log registra todas estas peticiones entrelazadas. La primera solución que podemos plantear es separar

esas entradas por dirección IP de origen, que, en principio, identifica máquinas diferentes y, por lo tanto, usuarios distintos. Un segundo refinamiento consiste en pensar que una misma persona puede realizar diferentes sesiones a lo largo del día. Para ello debemos establecer un periodo de tiempo tal que dos peticiones consecutivas realizadas desde una misma dirección IP separadas por este intervalo se consideren sesiones distintas. Tradicionalmente se han barajado tiempos en el intervalo de 10 a 30 minutos. En España, la OJD (Oficina para la Justificación de la Difusión) utiliza 10 minutos como tiempo de corte para diferenciar usuarios [5]. Establecer este intervalo para separar sesiones nos resuelve de paso otro problema: la asignación dinámica de direcciones IP. Cuando un usuario se conecta desde su casa a nuestro sitio web, lo más normal (todavía) es que lo haga mediante un proveedor de Internet que le ha asignado una dirección IP temporal, válida mientras dura su conexión. Cuando este usuario se desconecta, esa dirección IP es asignada a un nuevo navegante, que podrá acceder o no a nuestro sitio web. Por tanto, no es cierto que una dirección IP identifique a una máquina, aunque el establecimiento de ese espacio de tiempo para separar sesiones casi garantiza que diferentes personas, incluso con la misma dirección IP, serán consideradas como diferentes sesiones.

Aún nos queda un problema para el que la solución no es sencilla: en numerosas organizaciones el acceso a Internet es controlado a través de un proxy. El efecto de esta situación, en lo que al registro de accesos a un sitio web se refiere, es que todos los accesos producidos desde esa organización aparecen como realizados desde la misma dirección IP (la del proxy). Sólo podemos diferenciar a los usuarios que acceden a un sitio web mediante la utilización de cookies, que, a su vez, presentan otros problemas, como su mala imagen, y el que unos usuarios las acepten y otros no, con lo que los registros no serán en absoluto homogéneos. La solución más recomendable, desde nuestro punto de vista, es tratar estas direcciones IP como un único usuario, eso sí, con unas características especiales por su volumen de accesos y el tiempo que se conecta a nuestros servidores.

A medida que obtenemos datos derivados y agrupados de los accesos, se hace necesario estructurarlos de una manera adecuada. En este sentido, existen propuestas sobre el uso de DataWarehouses [6] adaptados a la información proveniente de sitios web [7] [8] que pueden ser utilizadas. Normalmente hacen uso de la sesión como elemento central, a partir del cual organizan el resto de los datos.

## 5 Las técnicas de minería de datos

Las técnicas de aprendizaje automático se han dividido tradicionalmente en supervisadas y no supervisadas [9]. Para las primeras se dispone de conjuntos de ejemplos que incluyen un valor que el sistema debe ser capaz de descubrir: son los sistemas de clasificación automática o predicción. El mecanismo de funcionamiento es siempre similar: se entrena al sistema con un conjunto de ejemplos que incluyen el valor a descubrir. Como resultado del proceso se obtiene un sistema capaz de "adivinar" ese valor cuando se le introducen datos, hasta ese momento desconocidos, Las técnicas más extendidas son, sin duda, los árboles de decisión [10, 11] para sistemas de clasificación (el valor a descubrir es discreto y representa la clase a la que

pertenece un objeto) y las redes neuronales artificiales, y dentro de ellas los perceptrones multicapa [12], para predicción de valores continuos[1].

Un segundo tipo de técnicas de aprendizaje automático son las no supervisadas. En este caso disponemos de un conjunto de datos, y lo que perseguimos es encontrar relaciones entre los mismos, patrones habituales de comportamiento, pero que son completamente desconocidos antes del análisis. La técnica no supervisada más conocida es el agrupamiento automático [13]. Estas técnicas detectan grupos de datos que se parecen entre sí, y que, a su vez, son suficientemente diferentes de otros grupos de datos. Existen diferentes algoritmos desarrollados, aunque el más extendido es el algoritmo de las K-medias, tanto en su versión más tradicional, como en su versión borrosa.

Para poder hacer uso de cualquiera de estas técnicas, tanto supervisadas como no supervisadas es necesario disponer de un gran volumen de datos, pero eso es precisamente de lo que más tenemos en un servidor web. La siguiente subsección presenta cuáles de estas técnicas son aplicables en el análisis de los datos recogidos en un servidor web y cómo pueden aportar información realmente valiosa a los informes estadísticos clásicos.

### 5.1 Técnicas de minería de datos aplicables al análisis de sitios web

En un primer momento, el tipo de técnicas de aprendizaje automático que vamos a utilizar son las no supervisadas. Trataremos de reconocer patrones habituales de comportamiento, que sean aplicables al conjunto de los visitantes del sitio web, para lo que podemos emplear diferentes técnicas. Un ejemplo concreto de aplicación real de todas ellas puede verse en [14].

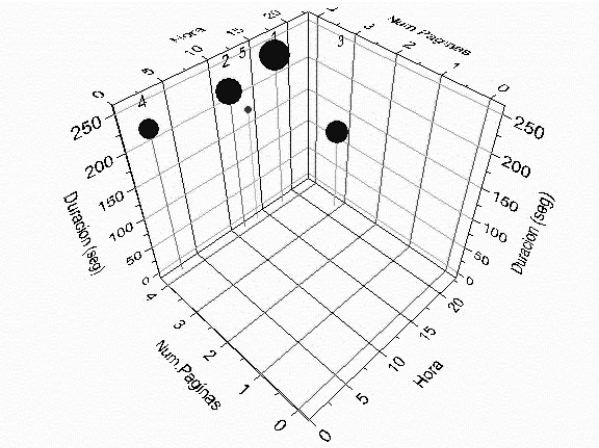
#### 5.1.1 Agrupamiento automático

En todo sitio web en el que se analicen las sesiones hay una serie de parámetros que definen cada una de éstas: el tipo de día en el que se realiza la conexión (laborable o festivo), la hora, el tiempo que el usuario se conecta a nuestro servidor y el número de páginas que visita. Como podemos ver, estos datos son completamente genéricos y aplicables a cualquier servidor web. ¿Para qué nos sirve el agrupamiento automático en este caso? Para establecer grupos de individuos similares que acceden a las páginas del sitio web. En las técnicas de agrupamiento automático, cada uno de estos grupos se va a identificar por un individuo prototipo. Para cada sesión, se establece el conjunto al que pertenece, de los generados automáticamente. En el caso de utilizar técnicas de agrupamiento borroso, cada sesión tendrá un cierto grado de pertenencia a cada uno de los conjuntos. La cardinalidad de cada conjunto nos determina el tamaño relativo del mismo.

La figura 2 muestra un ejemplo de agrupación automática que ha generado cinco prototipos. Como se puede apreciar, el sitio web analizado apenas tiene accesos en día

[1] Esto no quiere decir que no se haga uso de perceptrones multicapa para el desarrollo de sistemas de clasificación automática. Simplemente que se trata de una de las técnicas más extendidas para predicción.

festivo, casi todos los usuarios ven un promedio de 4 páginas en unos tres minutos, y la mayoría de los visitantes accede por la tarde. Esta información debe servir para replantear el diseño del sitio web y hacer, por ejemplo, que toda la información se encuentre a no más de 4 clics de ratón desde la página principal.




SEGMENTACIÓN DE SESIONES DE USUARIOS

| Grupo | Tipo día | Hora | Núm. páginas | Duración | Vol. de usuarios (%) |
|---|---|---|---|---|---|
| 3 | Laborable | 17 | 4,00 | 03m 57s | 28,96 |
| 4 | Laborable | 11 | 4,00 | 03m 41s | 25,30 |
| 5 | Laborable | 21 | 3,00 | 02m 13s | 20,60 |
| 2 | Laborable | 2 | 4,00 | 03m 48s | 18,76 |
| 1 | Festivo | 13 | 4,00 | 03m 06s | 6,38 |

**Fig. 2.** Agrupamiento automático de visitantes de un sitio web y prototipos que definen cada uno de los grupos generados, ordenados por su cardinalidad.

### 5.1.2 Análisis de asociaciones

Otra técnica no supervisada que puede ser utilizada en el análisis del comportamiento de los visitantes de un sitio web es el análisis de asociaciones [15], conocido vulgarmente como análisis del carro de la compra. Este nombre viene de su utilización más habitual: la detección de patrones de compra en grandes superficies. Para detectar estos patrones se analizan los tickets de venta, tratando de localizar relaciones del tipo, "los clientes que compran el producto A también compran el producto B". En muchos casos estas relaciones son obvias: mantequilla y galletas o cerveza y aceitunas; sin embargo, en ocasiones aparecen relaciones inesperadas (un ejemplo clásico es la relación aparecida en un hipermercado francés entre la compra de pañales y cerveza en fin de semana). En el caso de un sitio web, el análisis de las páginas visitadas en cada sesión nos permitirá conocer qué páginas suelen ser vistas conjuntamente por los visitantes. Nuevamente habrá relaciones que apenas aportarán información, como las que hay entre las páginas principal, y cualquiera que “cuelgue” de ella. Pero las relaciones entre páginas que no tienen un enlace directo sí aportan

una información real sobre el tipo de uso que hacen los usuarios del sitio web. El algoritmo más ampliamente utilizado es “apriori”, descrito con detalle en [15]. Ejemplos de información obtenida por el sistema pueden verse en la figura 3

```
El 40% de clientes/usuarios que accedieron la página
web con URL /entidad/productos/producto1.html, también
accedieron a /entidad/informacion/tema2.html;

El 30% de clientes/usuarios que accedieron a
/entidad/anuncio/oferta-especial.html, efectuaron un
pedido interactivo en /entidad/productos/producto1.
```

**Fig. 3.** Ejemplos de reglas de asociación

### 5.1.3 Análisis de secuencias

Por último, no debemos olvidar que el análisis del carro de la compra no tiene en cuenta si primero se compra el producto A y luego el B o al revés, ya que esa información no está disponible en el ticket de compra. Nosotros sí podemos hacer uso de esa distribución temporal de los accesos, distinguiendo si en la relación encontrada entre dos o más páginas, hay una relación de orden o no.

Para esto es posible hacer uso de algoritmos apropiados [16] o modificar el propio algoritmo a priori, para que se establezcan relaciones entre productos que mantengan una relación temporal. Para la aplicación concreta de análisis de un sitio web, cualquiera de las dos es aceptable.

## 5.2 Nuevas propuestas

Aunque las técnicas no supervisadas son empleadas normalmente en procesos de análisis de comportamiento de usuarios en sitios web, es posible también utilizar técnicas supervisadas. De este modo, se pueden utilizar algoritmos de predicción para tratar de establecer pautas de crecimiento de usuarios en nuestro sistema, de solicitudes realizadas a un servicio a medida que éste va siendo conocido por los visitantes, etc. En este caso, las técnicas supervisadas (por ejemplo, perceptrones multicapa) aprenderán de los datos almacenados sobre el comportamiento de los usuarios y aportarán una información fundamental para el correcto dimensionamiento, tanto de la infraestructura tecnológica, como del personal que deba atender el sistema, en su caso.

Por otro lado, una vez que se establecen grupos de usuarios en función de su comportamiento, haciendo uso de técnicas de agrupamiento automático, puede ser muy interesante generar las reglas que permiten clasificar nuevos visitantes en uno u otro grupo, de manera que esta clasificación pueda realizarse on-line. Para esto es posible hacer uso de árboles de decisión o de medidas de distancia respecto a los prototipos de los grupos generados.

A partir de la clasificación on-line de los usuarios, es posible ofrecer caminos adaptados a su perfil o contenidos específicos del grupo al que aparentemente pertenezcan. En la actualidad se hace uso de sistemas estadísticos para conseguir este mismo fin en sitios web de comercio electrónico.

## 6 Resumen y conclusiones

Los sitios web registran una cantidad ingente de datos sobre los accesos de los usuarios a sus páginas. Tradicionalmente, esos datos son analizados mediante técnicas estadísticas, pero su volumen hace que técnicas propias de la minería de datos puedan aportar información valiosa. Para poder aplicar estas técnicas es necesario un preprocesado de los datos, de manera que sea posible trabajar sobre las páginas solicitadas por los visitantes en cada una de las sesiones.

Las técnicas de reconocimiento de patrones, como el agrupamiento automático o el análisis de secuencias aportan un primer conjunto de algoritmos útiles. Por su parte, las técnicas supervisadas, fundamentalmente predicción y clasificación automática, no suelen ser utilizadas en este tipo de sistemas. Este trabajo sugiere cómo puede abordarse su uso.

Los autores de este artículo trabajan en la realización de aplicaciones reales que hacen uso de técnicas de minería de datos para el análisis de sitios web desde hace más de tres años. A pesar de que estas técnicas ofrecen una información que no puede ser descubierta fácilmente haciendo uso de técnicas más tradicionales como la estadística, aprecian que se trata de una tecnología que el usuario final acepta con dificultad. El motivo aparente tiene que ver con las expectativas del usuario cuando recibe un informe de un sitio web: estadísticas básicas del acceso al mismo. Para realizar una introducción real de una tecnología más avanzada, se hace necesario crear la necesidad, aportando ejemplos contundentes de las ventajas obtenidas en términos de una mejor información y conocimiento de los visitantes del sitio web. Hasta ese momento, debemos seguir trabajando en la localización de nuevas técnicas aplicables a este entorno y que permitan poner en el mercado real los algoritmos obtenidos como resultado de los proyectos de investigación en minería de datos.

## 7 Referencias


[1] CRISP-DM consortium CRISP 1.0 Process and User Guide. In http://www.crisp-dm.org/CRISPWP-0800.pdf

[2] Wirth R. and Hipp J., "CRISP-DM: towards a standard process model for data mining", Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD 2000), pp. 29-39. 11-13 April 2000; Manchester, UK

[3] Villena, J., González, J.C., Barceló, E. y Velasco, J.R., “Minería de uso de la web mediante huellas y sesiones”, IBERAMIA 2002, Sevilla, España.

[4] R. Kimball and R. Merz. The Data Webhouse Toolkit. John Wiley and Sons, Inc., 2000.

[5] OJD. Reglamento de trabajo para el control de medios electrónicos de comunicación, en http://www.ojd.es/Aregla/mec/home.htm

[6] Kimball, R y Ross, M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, John Wiley & Sons, 2002.
[7] Mattison, R., Web Warehousing and Knowledge Management. McGraw-Hill, 1999
[8] E. Barceló, J.Villena, J.R.Velasco; Desarrollo de un Sistema de Minería de Uso de la Web en Tiempo Real. CAEPIA 2001.
[9] Berthold, M. y Hand, D., Intelligent Data Analysis: An introduction. Springer, 1999
[10] Quinlan, J.R., "Learning efficient clasification procedures and their application to end chess games", en Michalsky, R.S., Carbonell, J.G. y Mitchell, T.M. eds. Machine Learning: An Artificial Intelligence Approach, Springer-Verlag, Berlin, 1984.
[11] Quinlan, J.R. C4.5: Programs For Machine Learning Morgan Kaufmann, San Mateo, CA, 1993.
[12] Lippmann, R.P., " An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, April 1987, pp 4-22.
[13] Bezdek, J.C., Pattern Recognition with Fuzzy Objective Function Algorithms., New York: Plenum, 1981.
[14] DAEDALUS, Lawerinto Miner, Descripción de producto, en http://www.daedalus.es/ DocumentacionLwMiner.asp, 2001.
[15] Agrawal, R., Imielinski, T. y Swami. A., "Mining association rules between sets of items in large databases". SIGMOD-Record. vol.22, no.2; June 1993; p.207-216.
[16] Manilla, H., Toivonen, H. & Verkamo, A. 1995. Discovering Frequent Episodes in Sequences. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, 210-215. AAAI Press.