tecnimap 2002
www.tecnimap.com
VII JORNADAS SOBRE TECNOLOGÍAS DE LA INFORMACIÓN PARA LA MODERNIZACIÓN DE LAS ADMINISTRACIONES PÚBLICAS
A CORUÑA
15, 16, 17 y 18 de octubre de 2002
Palacio de Congresos y de la Ópera
Ponencias

Uso de la Minería Web para mejorar los servicios al ciudadano

Julio Villena Román

Fernando García Vegas

José Carlos González Cristóbal

1. Introducción

Las tecnologías de la información y comunicación forman un instrumento clave para el desarrollo local. Promoviendo el acceso universal a servicios y aplicaciones basados en las nuevas tecnologías de la información y comunicación, se puede mejorar la calidad de vida de los ciudadanos, utilizando la tecnología como pilar que soporte los desafíos de la cohesión socioeconómica, ciudadanía, inclusión social, democratización y desarrollo sostenible. Y en este sentido, las autoridades locales y regionales juegan un papel fundamental en el desarrollo de la sociedad de la información.

Las tecnologías digitales permiten un acceso y una reutilización más fáciles del acervo de información que posee el sector público. La administración electrónica podría transformar la organización tradicional del sector público y proporcionar unos servicios más rápidos y más sensibles a las necesidades. Asimismo, puede aumentar la eficiencia, rebajar costes, lograr una mayor transparencia y simplificar los trámites administrativos de las empresas y los ciudadanos. El acceso electrónico supone también una aportación destacada para acelerar la transición a la sociedad de la información estimulando los servicios de Internet más interesantes para los ciudadanos.

El proceso de modernización de las Administraciones públicas es tan deseable como imparable y cada vez más el ciudadano se encuentra y puede acceder a una creciente cantidad de servicios ofrecidos utilizando Internet.

Hoy en día, todos y cada uno de los ministerios y una inmensa mayoría de organismos de las diferentes administraciones públicas mantienen uno o varios sitios web, los cuales son cada día más sofisticados y ofrecen una interacción más detallada y compleja con los ciudadanos. Algunos de los servicios ofrecidos han tenido un enorme éxito y quizá uno de los mejores ejemplos sea el programa PADRE y la posibilidad que ofrece la Agencia Tributaria de presentar la declaración de la renta a través de Internet.

En este entorno, la integración de técnicas de CRM orientadas a la gestión de las interacciones con los ciudadanos - visitantes de los sitios web - y la explotación de las mismas se convierte en uno de los puntos esenciales del proceso. Debido a la naturaleza de los datos manejados, la información que se obtiene a través de la interacción de los usuarios con los portales de la administración se va a convertir en una de las más valiosas fuentes a la hora de evaluar y mejorar la calidad del servicio ofrecido. La correcta adquisición y tratamiento de esta información por tanto va a ser un asunto clave.

De esta manera, el uso de técnicas de Minería de Datos aplicadas al análisis del uso de los sitios web (más conocido por su nombre en inglés, web mining) junto con una correcta técnica de adquisición de datos se convierte en este marco en una herramienta indispensable.

En este documento se pretenden presentar los problemas asociados y las soluciones que habría que implementar en una herramienta de Minería Web para que los resultados que ofreciera fueran útiles y veraces. A continuación se detallará una arquitectura propuesta de una herramienta que cumpliera todos los requisitos. Finalmente se detallarán las ventajas que tendría una herramienta de este tipo para un gestor de un sitio web en general y para los portales de la Administración en particular.

2. Minería Web

La minería de datos en la web (web mining) es la aplicación de la minería de datos (data mining) sobre información existente en la web (Internet/Intranet) complementada con otro tipo de información, empleando técnicas de extracción de patrones interesantes y potencialmente útiles para descubrir correlaciones y tendencias significativas para dicha organización [6].

La clasificación más típica distingue básicamente entre tres dominios de aplicación: **minería de contenidos en la web** (extracción de conocimiento sobre el contenido de los documentos o sus descripciones, incluye la minería de documentos de texto o text mining), **minería de estructura de la web** (inferencia de conocimiento de la organización de la web y de los enlaces entre referencias y referentes en la web) y, finalmente, **minería de uso de la web**, como el proceso de extracción de patrones interesantes de la información del tráfico en la web. Algunos autores [6] no incluyen la minería de estructura en su clasificación.

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.

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 tres problemas fundamentales.

En primer lugar, 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. 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. Este problema puede ser subsanado utilizando otra forma de recolección de los accesos, empleando las llamadas huellas.

En segundo lugar, 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. La importancia de este tipo de sistemas es cada día más importante y se espera que esta importancia no haga más que crecer en años venideros. 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. De nuevo el uso de huellas proporciona una herramienta para realizar fácilmente esta distinción

El tercer problema son las métricas (unidades) que se emplean tradicionalmente en los análisis, que son los accesos a páginas. 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.

A continuación se exponen los fundamentos del método de las huellas y la definición y la utilidad de adoptar la sesión de usuario como métrica del uso, calculada a partir de los accesos.

3. Registro del tráfico mediante huellas

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. La información que se almacena en el log depende de cada servidor y su configuración particular, pero 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/devueltos, la identificación del cliente (User-Agent), cookies y la página de procedencia (Referrer).

Como ya hemos comentado, el uso del log plantea un serio problema al no quedar en el registrados todos los accesos que realmente se producen, 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.

La solución alternativa que se propone en este documento 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. Usando técnicas que eviten que este elemento pueda ser introducido en las caches se puede garantizar que no se pierde ninguna petición de usuario.

4. 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.

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.

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.

Además este tipo de distinción permite realizar análisis y clasificaciones de los usuarios basados en el **comportamiento de cada sesión** permitiendo identificar por ejemplo las diferentes pautas de navegación de los diferentes tipos de usuarios.

5. Arquitectura de un sistema real

Aquí se presenta la descripción de un sistema [7] sencillo de usar que permite obtener informes automáticos de acceso y uso de un sitio web, que van desde los estadísticos más relevantes a informes avanzados de minería de datos como segmentación de visitantes, asociaciones entre páginas y secuencias de páginas vistas. La principal diferencia entre este sistema y una herramienta convencional de logs es precisamente que no requiere acceder a estos ficheros, sino que todos los datos necesarios se obtienen en tiempo real de un sistema de bases de datos o webmart, en el que los datos han sido almacenados mediante una huella [3].

La característica más relevante de este sistema de minería de uso de la web en tiempo real es que el cálculo de todos los informes y estadísticas se desarrolla en función del concepto de **sesión de usuario** [8].

El sistema desarrollado se compone de cuatro módulos principales. El **webmart** es el almacén central de toda la información sobre el uso del sitio web empleada por los demás módulos del sistema. **El módulo de carga** es el encargado de insertar nueva información en el webmart. El tratamiento, procesamiento y análisis de la información se realizan a través del **módulo de análisis** de datos. Por último, el **módulo de visualización** se encarga de la presentación dinámica de informes y la navegación por ellos.

El diseño hace hincapié en la independencia y complementariedad de los módulos. Cada módulo tiene una labor perfectamente definida que no se superpone con la de otros módulos, siendo actualizable por nuevas versiones con nuevas o distintas funcionalidades sin afectar al resto. Así el sistema es más adaptable a diferentes escenarios de uso (por ejemplo, el webmart se adaptaría a la tecnología de acceso de datos existente, sin afectar al resto), más escalable (si un módulo se convierte en cuello de botella se podría replicar, sin afectar al resto) y mantenible (los errores están más localizados).

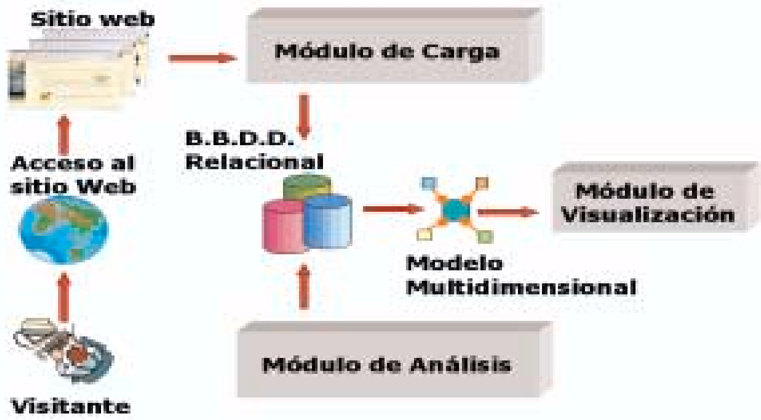


Figura 1. Esquema funcional del sistema de minería de uso de la web en tiempo real

El webmart

El webmart [8] es el almacén de datos que contiene toda la información. Se denomina así por analogía con un datamart con datos de uso de un sitio web. Se compone de tres conjuntos lógicos de tablas relacionales que contienen diferentes tipos de información (Figura 2).

El primer conjunto son las tablas de **mediciones de tráfico**, que contiene, como su nombre indica, información sobre las mediciones de tráfico sobre un determinado sitio web. Se compone de dos tablas: **la tabla trasformada** que contiene toda la información tratada y filtrada de la tabla operacional, que es la que registra y mantiene los accesos (hits) instantáneos a las páginas que son insertados por el módulo de carga, y la **tabla de sesiones**, que contiene las diferentes sesiones que agrupan todos los accesos a las páginas realizados por un mismo visitante, y son calculados en el primer proceso del módulo de análisis.

El segundo conjunto lógico de tablas son las tablas de **estadísticas básicas**, que resumen y agregan la información de las tablas anteriores para permitir un acceso rápido a diferentes estadísticas sobre el tráfico en un determinado periodo de tiempo (un día, una semana, una quincena, y un mes), entre ellas: las páginas más accedidas, las páginas más empleadas como entrada y como salida en una sesión, máquinas de origen, páginas de procedencia, distribución de visitas por hora del día, por días, por dominios, tecnologías de acceso del cliente, etc.

El tercer y último conjunto de tablas son las tablas de **minería de datos**, que contienen los resultados del procesamiento de diferentes algoritmos inteligentes de procesamiento del tráfico de un sitio web que pretenden encontrar patrones para modelar el comportamiento de los visitantes (asociaciones de páginas vistas [1], secuencias entre páginas [2] y agrupamiento de visitantes [4]) y que constituyen el estado del arte en minería de uso de la web.

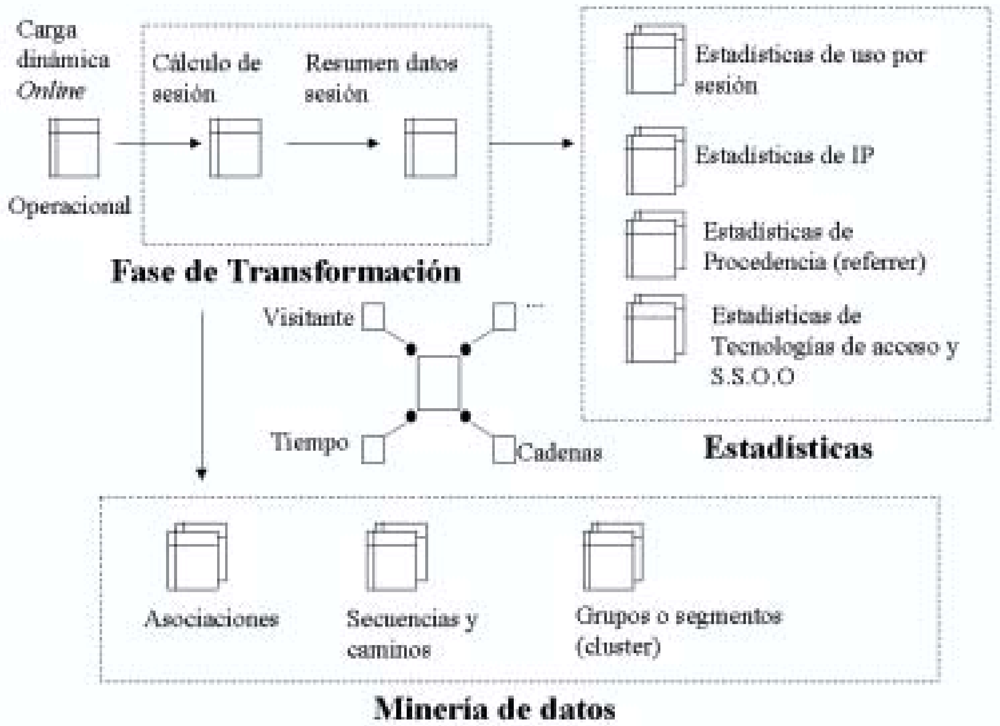


Figura 2. Estructura lógica del webmart

Módulo de carga

Este módulo es el encargado de realizar la inserción de nueva información en el webmart. Esta carga se puede realizar mediante dos mecanismos. El primero de ellos (y el más deseable) es mediante huellas en las páginas web. Como alternativa se puede cargar el webmart a partir del propio log del servidor, pero en este caso se deben tener en cuenta los problemas asociados al uso de esta fuente de información.

Módulo de análisis

El módulo de análisis consta de varios componentes: el módulo de cálculo de sesiones, el módulo de obtención de estadísticas básicas y los módulos de algoritmos de minería de datos sobre el uso de la web. Todos ellos son servicios automáticos que se ejecutan periódicamente cada cierto tiempo, completando tablas del webmart.

DISTRIBUCIÓN DE ACCESOS POR DÍAS

Fecha	Accesos (clics)	Sesiones	Páginas promedio	Duración promedio
20/01/2002	32	9	3,56	02m 11s
21/01/2002	194	35	5,54	04m 34s
22/01/2002	189	50	3,78	03m 04s
23/01/2002	596	82	7,27	06m 33s
24/01/2002	755	75	10,07	12m 07s
25/01/2002	575	50	11,50	23m 12s
26/01/2002	86	21	4,10	05m 10s
27/01/2002	91	23	3,96	02m 43s
28/01/2002	238	51	4,67	04m 21s
29/01/2002	337	54	6,24	04m 10s
30/01/2002	299	57	5,25	08m 16s
Totales	Total Accesos	Total Sesiones	Páginas promedio	Duración promedio
	3392	507	6,69	07m 57s

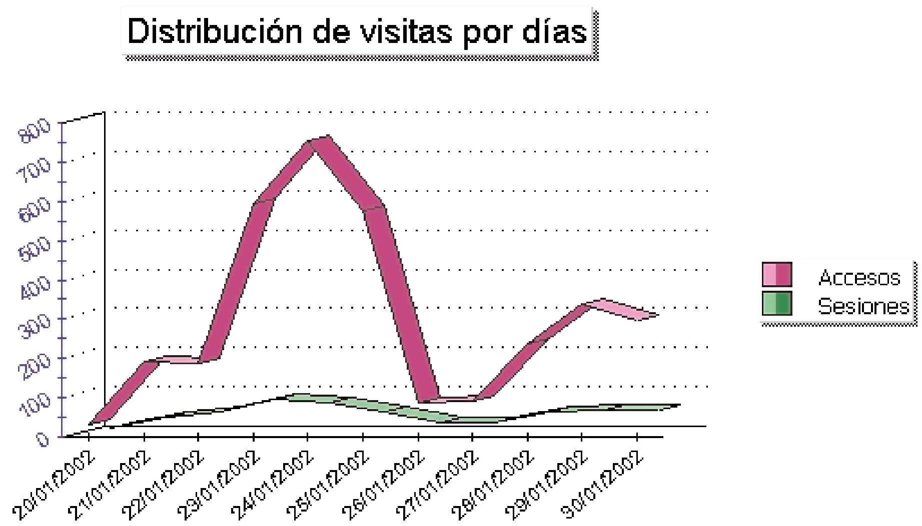


Figura 3. Resumen de distribución de accesos por días

Módulo de visualización

El módulo de visualización muestra de forma más comprensible los informes de resultados de los diferentes módulos. La implementación actual hace uso de páginas dinámicas que muestran en forma de tablas y gráficos las estadísticas (Figura 3) e informes de minería de datos (Figuras 4, 5 y 6). La generación de estos informes en función de un intervalo temporal, ya sea de horas, días, semanas, meses se realiza on-line. Además, es una características fundamental de este sistema el hecho de poder generar informes en tiempo real consultando por conceptos tales como los dominios de acceso, los visitantes, la tecnología, páginas, asociaciones, secuencias más frecuentes, etc.

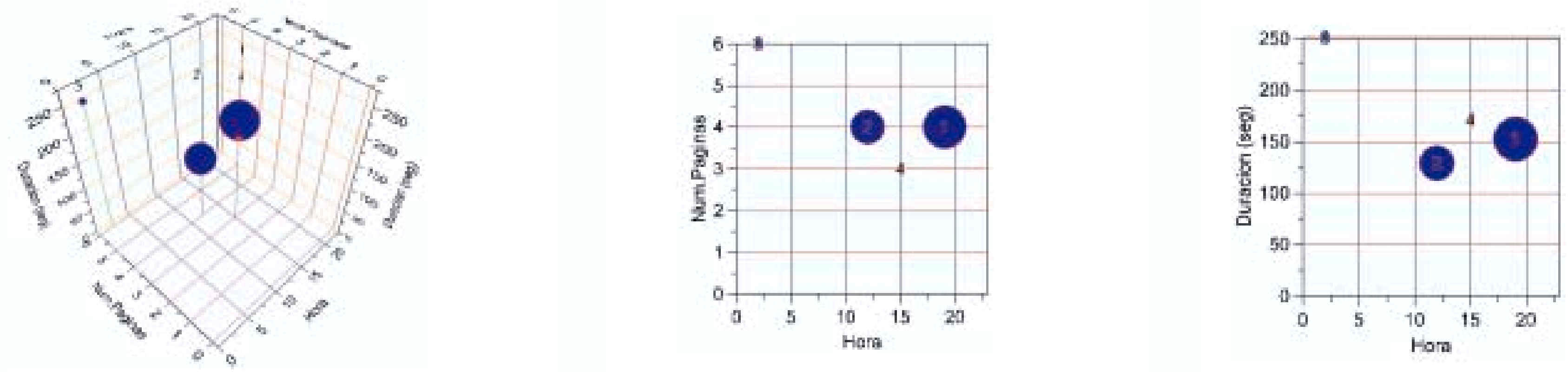
6. Algoritmos de Minería de datos

La parte fundamental de este sistema es el uso de algoritmos inteligentes de búsqueda de patrones de comportamiento para modelar a los visitantes. Los algoritmos utilizados detectan asociaciones de páginas vistas, secuencias entre páginas y caminos seguidos y clasifican a los visitantes en grupos homogéneos según unos ciertos parámetros.

Agrupamiento de sesiones de visitantes

El sistema agrega automáticamente a los visitantes en función de algunos parámetros válidos para el análisis de cualquier sitio web. Los parámetros o variables utilizados para agrupar a los visitantes en grupos homogéneos son cuatro: el tipo de día de la semana, la hora de comienzo de la visita, la duración de la visita y el número de páginas vistas.

Con estos cuatro parámetros se ha usado el algoritmo Fuzzy-C-Means [4] para el agrupamiento automático de las visitas. Los mecanismos aplicados [4] [10] distribuyen a individuos de comportamientos similares en grupos homogéneos. A cada grupo se le asigna un elemento prototipo o ideal que refleja las características básicas de los componentes del grupo. El resultado del proceso de segmentación es un conjunto de prototipos, y una distribución volumétrica del conjunto de sesiones, calculado como la suma de los grados de pertenencia de los elementos, sesiones, a los prototipos encontrados, entre los diferentes grupos. El procedimiento de cálculo obtiene el número de clases en las que se agrupan idealmente los visitantes o sesiones. En la Figura 7 se muestra el resultado de la segmentación en un intervalo de 15 días para el sitio web de estudio.



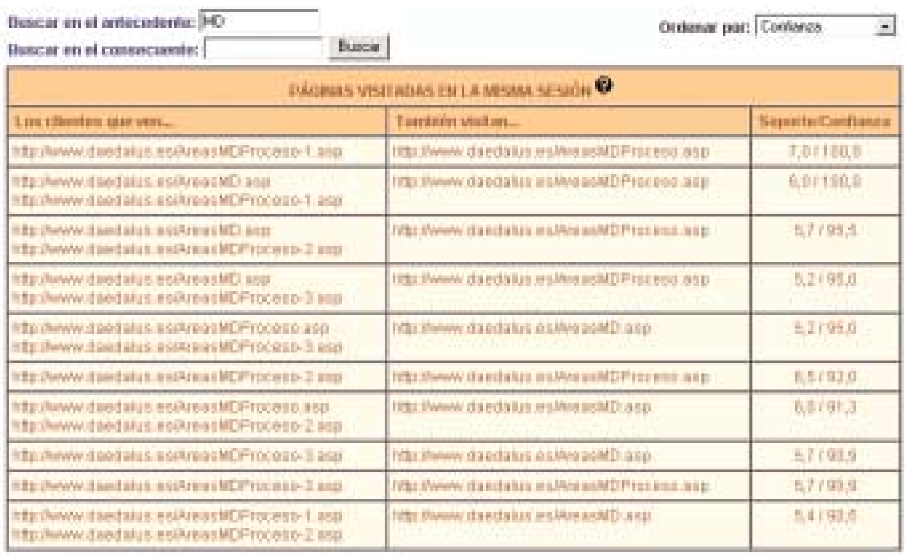
SEGMENTACIÓN DE SESIONES DE USUARIOS					
Grupo	Tipo día	Hora	Núm. páginas	Duración	Vol. de usuarios (%)
1	Laborable	19	4,00	02m 33s	47,54
3	Laborable	12	4,00	02m 09s	37,60
4	Laborable	2	6,00	04m 11s	8,71
2	Festivo	15	3,00	02m 51s	6,15

Figura 4. Ejemplo de segmentación de sesiones de usuario en un sitio web

Reglas asociativas entre páginas

Los mecanismos de búsqueda de asociaciones de páginas (visitadas en una sesión) permiten determinar qué conjuntos de páginas aparecen agrupadas (independientemente de su orden) de forma reiterada en un número significativo de sesiones. El algoritmo utilizado es el denominado apriori [1].

Las asociaciones se expresan normalmente en forma de reglas como la siguiente: en las sesiones en que se visitan las páginas A y B, también se accede a la página C. Cada asociación queda cuantificada con dos parámetros: soporte (proporción de sesiones en las que se da la primera condición, visita de A y B) y confianza (proporción de las anteriores, en las que también se cumple la segunda condición, acceso a C). De esta forma, la regla anterior se podría interpretar como: el X% (soporte) de los visitantes acceden a las páginas A y B, y de ellos, el Y% (confianza) también acceden a la página C.


Buscar en el antecedente: MD
Buscar en el consecuente:
Buscar
Ordenar por: Confianza

PÁGINAS VISITADAS EN LA MISMA SESIÓN		
Los clientes que ven...	También visitan...	Soporte/Confianza
http://www.daedalus.es/AreasMDProceso-1.asp	http://www.daedalus.es/AreasMDProceso.asp	7,0 / 100,0
http://www.daedalus.es/AreasMD.asp http://www.daedalus.es/AreasMDProceso-1.asp	http://www.daedalus.es/AreasMDProceso.asp	6,0 / 100,0
http://www.daedalus.es/AreasMD.asp http://www.daedalus.es/AreasMDProceso-2.asp	http://www.daedalus.es/AreasMDProceso.asp	5,7 / 95,4
http://www.daedalus.es/AreasMD.asp http://www.daedalus.es/AreasMDProceso-3.asp	http://www.daedalus.es/AreasMDProceso.asp	5,2 / 95,0
http://www.daedalus.es/AreasMDProceso.asp http://www.daedalus.es/AreasMDProceso-3.asp	http://www.daedalus.es/AreasMD.asp	5,2 / 95,0
http://www.daedalus.es/AreasMDProceso-2.asp	http://www.daedalus.es/AreasMDProceso.asp	6,5 / 93,0
http://www.daedalus.es/AreasMDProceso.asp http://www.daedalus.es/AreasMDProceso-2.asp	http://www.daedalus.es/AreasMD.asp	6,8 / 91,3
http://www.daedalus.es/AreasMDProceso-3.asp	http://www.daedalus.es/AreasMD.asp	5,7 / 93,9
http://www.daedalus.es/AreasMDProceso-3.asp	http://www.daedalus.es/AreasMDProceso.asp	5,7 / 93,8
http://www.daedalus.es/AreasMDProceso-1.asp http://www.daedalus.es/AreasMDProceso-2.asp	http://www.daedalus.es/AreasMD.asp	5,4 / 93,6

Figura 5. Extracto de las asociaciones entre páginas en un periodo de 15 días

Secuencias y caminos recorridos

Los mecanismos de búsqueda de patrones secuenciales permiten determinar qué secuencias ordenadas (y correlativas) de páginas aparecen de forma reiterada en un número significativo de sesiones. Las secuencias se expresan normalmente en forma de reglas como la siguiente: en las sesiones en que se visita la página A seguida de la B, se accede después a la página C. Cada secuencia queda cuantificada con dos parámetros: soporte (proporción de sesiones en las que se da la primera condición, visita de A seguida de B) y confianza (proporción de las anteriores, en las que también se cumple la segunda condición, acceso subsiguiente a C). En consecuencia se crearán reglas como: el X% de los visitantes acceden a la página A seguida de la B, y de ellos, el Y% acceden inmediatamente después a la página C.

La diferencia entre la búsqueda de patrones secuenciales y de asociaciones es que las primeras conservan el orden en el que los clientes visitan las páginas, mientras que para las asociaciones es indiferente el orden el que éstas fueron visitadas.

7. Minería de uso web en el marco del CRM para la Administración

Todos los avances tecnológicos posibilitan que las personas puedan acceder y conocer con facilidad, diversas alternativas de solución a sus necesidades específicas, provocando que sus exigencias o demandas sean más personalizadas, específicas y concretas, provocando también que reclamen sus derechos a la administración pública mayor oportunidad y calidad, para así obtener un mayor valor del servicio público.

El desafío no está en la incorporación misma de la tecnología en términos de su aplicación, sino que es un proceso a desarrollar en forma integral y que va desde la administración interna, pasando por la participación, el rediseño de las culturas organizacionales, hasta la ope-

CAMINOS HABITUALMENTE SEGUIDOS		
Páginas vistas en orden	Soporte/Confianza	Nº Casos
Antecedente: http://www.gsi.dit.upm.es/~gfe/spanglish/indice.html http://www.gsi.dit.upm.es/~gfe/spanglish/titulo.html Consecuente: http://www.gsi.dit.upm.es/~gfe/spanglish/inicial.html	4,28 / 77,4	62
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/ http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html	4,91 / 67,1	71
Antecedente: http://www.gsi.dit.upm.es/~gfe/spanglish/ http://www.gsi.dit.upm.es/~gfe/spanglish/titulo.html Consecuente: http://www.gsi.dit.upm.es/~gfe/spanglish/indice.html	5,46 / 58,2	79
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/ http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/portada.html	4,91 / 63,8	71
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/ http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/blanca.html	4,91 / 63,8	71
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/ http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/blanca.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/portada.html	3,18 / 95,7	46
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/portada.html	4,28 / 71,0	62
Antecedente: http://www.gsi.dit.upm.es/~gfe/ssii/ http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/index.html http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/blanca.html Consecuente: http://www.gsi.dit.upm.es/~gfe/ssii/rcsi/portada.html	3,11 / 95,6	45

Figura 6. Secuencias de páginas vistas en orden en un periodo quincena

ración de una estrategia de trabajo en red que supere las intervenciones sectoriales. Es el inicio del Gobierno local en línea.

En este marco la utilización de las técnicas CRM para mejorar la calidad del servicio ofrecido a los ciudadanos por la Administración y conseguir una gestión más eficaz en costes, velocidad y satisfacción del usuario es más una exigencia que una opción. Pero para poder aplicar estas técnicas es necesario que los datos de partida sean completamente fiables, ya que sólo a partir de una buena información es como se puede construir un eficaz sistema de gestión a todos los niveles.

Si unimos esto al hecho innegable de que cada vez más y más gestiones se realizan usando tecnologías del mundo Internet, y muy especialmente gestiones remotas desde el domicilio o lugar de trabajo de los ciudadanos usando un navegador para acceder a servicios on-line, una buena herramienta de minería de uso web puede contribuir enormemente al resultado global de cualquier avance en este sentido, así como a la evaluación de su éxito de forma prácticamente instantánea.

La información recogida sobre las interacciones de los visitantes de un sitio web tiene un amplio espectro de usos y posibilidades, no solo para conocer las páginas y contenidos en los están interesados los usuarios y cómo y cuándo los visitan, sino también su aplicación para ayudarles, aconsejarles y guiarles en sus operaciones y actuaciones. Esto es de especial importancia para tres casos específicos, todos los cuales se aplican en el caso del acceso a sitios web o portales de las Administraciones Públicas.

En primer lugar, es muy importante ayudar a los usuarios cuando el público que visita las páginas tiene un perfil de usuario general sin ser un gran experto de los ordenadores, la navegación por Internet y las tecnologías de la información, para que no se sienta perdido por el simple hecho del empleo de la tecnología y renuncie a la visita. Dado que el público objetivo de un portal de la Administración tiene este perfil, sería necesario que todos los portales diseñados incluyeran consideraciones específicas de navegabilidad, usabilidad, sencillez de manejo y ayudas al usuario.

Por otra parte, es importante ayudar al usuario, sea éste o no un experto en la navegación por Internet, si el proceso que está realizando tiene cierta complejidad o la información que necesita manejar es abundante. En el caso de la e-administración, se incluyen procesos y gestiones que conllevan un grado de dificultad no despreciable, que es necesario simplificar lo máximo que sea posible. Además, la información recogida cubre un amplio abanico de temas y aspectos diferentes, desde información general sobre una ciudad hasta información específica de eventos (agenda local) o la posibilidad de ejecutar todo tipo de procesos y gestiones administrativas. Por ello cualquier tipo de ayuda, indicación o sugerencia al ciudadano de dónde encontrar lo que necesita y cómo hacerlo es extremadamente importante.

Por ultimo, por motivos de la propia eficiencia del impacto del sitio web. Cuando un portal es visitado por un gran número de personas (como puede ser el caso de un portal de un ayuntamiento, el de una administración autonómica o aún más el de un organismo nacional) y almacena gran cantidad de información de diverso tipo, es necesario que la información contenida esté bien estructurada y sea fácil acceder a ella y encontrar lo que se quiere.

Para todas estas necesidades una herramienta de minería de uso web proporciona los datos y análisis necesarios que van a permitir conocer con exactitud el uso real que se está haciendo de los recursos disponibles y que puede diferir de los objetivos que se propusieron en su elaboración. Además con este tipo de soluciones se puede optimizar fácilmente la navegación de los usuarios para hacerla lo más sencilla y rápida posible, permitiendo tanto mejorar la impresión que estos se llevan de los recursos ofrecidos, como ajustar los recursos disponibles para crear soluciones que garanticen el éxito de la gestión ofrecida.

Además, en la solución aquí presentada se aplican técnicas de minería de datos que permiten descubrir información sobre los hábitos de los usuarios que ayudarán a mejorar el servicio. Con el agrupamiento automático de los visitantes, se pueden diseñar estrategias proactivas para visitantes futuros personalizando los recursos ofertados en tiempo real. El análisis de asociaciones y secuencias suponen una herramienta indispensable a la hora de plantearse una posible mejora o rediseño del mismo, así como a la hora de localizar posibles fallos o errores en su creación.

8. Conclusiones

La correcta utilización y procesado de los datos obtenidos a partir de las visitas de los usuarios de un recurso web supone una gran ayuda a la hora de evaluar el resultado que este recurso está realmente obteniendo.

En caso de detectarse anomalías en el uso o simplemente si se desea mejorar la calidad de cualquier gestión realizada on-line, los datos obtenidos por una herramienta de Minería Web son el mejor punto de partida para acometer esta mejora.

Los beneficios que tendría la implantación de alguna herramienta de este tipo en la administración son aún más notables, debido básicamente a la cantidad y heterogeneidad de las visitas que sus portales y sitios web reciben.

El método de adquisición de datos es fundamental si se espera obtener resultados fiables y concluyentes. Los métodos tradicionales presentan unas carencias que deben ser subsanadas. La utilización de huellas puede arreglar todas las carencias que tiene es análisis de los logs del servidor.

Las métricas de uso basadas en cantidad de accesos (hits) que recibe un sitio web no son concluyentes a la hora de evaluar su éxito o fracaso. Las métricas basadas en sesiones de usuario son mucho más fiables.

El uso de algoritmos de Minería de Datos en estos sistemas ofrece un valor añadido muy importante, y supone una ayuda indispensable a la hora de analizar el comportamiento real de los visitantes.