LAS POSIBILIDADES OFRECIDAS POR EL "WEB MINING"

Álvaro Gómez Vieites
agomezvieites@gmail.com
Profesor de la Escuela de Negocios Caixanova

RESUMEN DE LA PONENCIA

Esta ponencia se centra en el análisis de las posibilidades ofrecidas por el "Web Mining", término que podríamos traducir por "minería de los datos registrados en el Website" de una organización.

A partir del análisis de los registros de conexión a un Website es posible determinar cuáles son las páginas y contenidos de dicho Website que tienen más éxito, evaluar los resultados de las campañas de promoción on-line (respondiendo a la pregunta de si es rentable la inversión en publicidad en determinados sitios dentro de Internet), estudiar fallos y deficiencias en las conexiones (porcentaje de sesiones, perdidas, transacciones incompletas…), etc., obteniendo una valiosa información que permitirá mejorar los contenidos y servicios incluidos en el Website.

Asimismo, la identificación de los visitantes a un Website permite crear una valiosa base de datos de marketing, que puede ser utilizada para ofrecer productos y servicios a medida, personalizar el proceso de comunicación y desarrollar otras técnicas incluidas dentro del Marketing "One-to-One".

También es posible definir el perfil de cada uno de los visitantes a partir del seguimiento de sus pasos dentro de un Website en sus distintas conexiones: qué páginas visita, cuánto tiempo le dedica a cada una de las secciones, qué productos y servicios busca dentro del catálogo de la empresa, etc. Con toda esta información los responsables del Website podrán adaptar el contenido y la información en función de cada uno de los perfiles generados.

1. EL POTENCIAL DE INTERNET COMO NUEVO MEDIO DE COMUNICACIÓN

Internet es un medio digital interactivo que permite desarrollar una comunicación directa y personalizada con cada cliente, sin limitaciones geográficas (puede ofrecer una cobertura global con la publicación de páginas Web) ni temporales. Además, a través de un mismo canal es posible realizar distintas interacciones con los clientes: publicidad e información preventa, configuración de pedidos, compras, servicios posventa, etc.

La naturaleza bidireccional de este canal permite desarrollar el concepto de personalización hasta sus últimas consecuencias:

- Presentación de contenidos totalmente adaptados a las necesidades de cada cliente: catálogos de productos Web, mensajes publicitarios y otros servicios.
- Posibilidad de desarrollar Websites adaptativos, con una estructuración de elementos (marcos, enlaces, botoneras) y un diseño (colores, tamaños del texto, resolución de las imágenes) que se pueden modificar de acuerdo con las preferencias manifestadas por los usuarios.
- Incorporación de sistemas de recomendación dentro del Website, que tienen en cuenta las características sociodemográficas, hábitos y perfiles de los clientes.
- Desarrollo de productos y servicios a medida: ordenadores, música, servicios de información, etc.

- ❖ Participación del cliente en la configuración del producto.
- ❖ Seguimiento de eventos clave en la vida del cliente: cumpleaños, aniversarios, sustitución de productos..., para poder anticiparse a sus necesidades.

Las empresas pueden utilizar Internet para mejorar la relación con sus clientes de varias formas: como plataforma para ofrecer información y contenidos, como nuevo canal de venta y distribución de sus productos, como medio de comunicación directa con los clientes (recurriendo a servicios como el correo electrónico o la telefonía IP) y como plataforma para proporcionar determinados servicios (como los servicios de asistencia posventa).

2. CONSTRUCCIÓN DEL WEBSITE CORPORATIVO

Dentro de Internet el Website se convierte en la delegación virtual de la organización, a través de la cual se podrá poner en contacto con distintos agentes: clientes, proveedores, empleados, público en general, etc.

La historia nos demuestra que no ha sido fácil explotar el potencial propio de cada nuevo medio de comunicación. Así, las primeras estaciones de radio comercial no eran poco más que periódicos con voz, mientras que los primeros programas de televisión eran programas de radio con imágenes. Por eso, hoy en día muchos Websites tratan de imitar a los catálogos en papel o a los programas de televisión, desaprovechando las características específicas de este nuevo medio digital.

Las claves del éxito en el diseño y construcción de un Website pasan por incluir contenidos y servicios que sean útiles para el usuario, que se encuentren organizados de forma clara y sencilla, aprovechando la interactividad para ofrecer un adecuado nivel de personalización, prestando especial atención a la facilidad de la navegación por los distintos contenidos, elementos y servicios incluidos en el Website, así como a la agilidad y rapidez en el acceso a las páginas Web.

En definitiva, se trataría de proporcionar al usuario de una manera amena y sencilla todo aquello que busca o que le pueda interesar dentro del Website de la empresa.

Podemos considerar que un proyecto de desarrollo de un Website corporativo consta de las siguientes etapas:

1. Definición del tipo de Website que se quiere construir y de los objetivos que con él se persiguen: tener presencia en Internet para informar de las actividades y de la historia de la organización, contribuyendo además a la mejora de su imagen; desarrollo de un catálogo electrónico de productos; creación de una comunidad virtual de usuarios; ofrecer nuevos servicios y soporte técnico a los clientes; venta directa de productos on-line a través del propio Website; etc.

2. Diseño y construcción de las páginas Web y de las bases de datos que constituyen el Website.

3. Puesta en marcha del servidor Web en Internet: se pueden barajar las alternativas de utilizar un servidor propio con una conexión permanente a la Red, aunque también se podría recurrir a una externalización del servicio a través del hosting (hospedaje de páginas Web) o del housing (ubicación de un servidor propio de la empresa en las instalaciones de un proveedor de acceso a Internet).

4. Promoción del Website: etapa fundamental para dar a conocer los servicios del Website a su público objetivo y generar tráfico, recurriendo para ello al registro en los buscadores, campañas de publicidad mediante anuncios en Internet como los banners, publicidad en medios tradicionales, marketing a través del correo electrónico, etc.

5. Medición de los resultados: control del tráfico recibido por el Website; seguimiento de las visitas; control de la efectividad de la publicidad en Internet; etc.

6. Mantenimiento y actualización del Website.

El Website de la empresa debe incorporar contenidos interesantes y útiles para sus visitantes, presentados de forma amena y atractiva, sacando el máximo partido de los hiperenlaces y de las características multimedia.

Además, estos contenidos deberían ser actualizados con frecuencia (dependiendo, lógicamente, de su naturaleza), manteniendo una misma línea editorial y una normativa interna para la publicación de los contenidos. Por otra parte, se deberían destacar las novedades y la fecha de última actualización del Website, para que los visitantes asiduos puedan dirigirse rápidamente a las secciones que han registrado cambios.

3. EXPLOTACIÓN DE LA INFORMACIÓN REGISTRADA EN INTERNET

A partir del análisis de los registros de conexión a un Website es posible determinar cuáles son las páginas y contenidos de dicho Website que tienen más éxito, evaluar los resultados de las campañas de promoción on-line (respondiendo a la pregunta de si es rentable la inversión en publicidad en determinados sitios dentro de Internet), estudiar fallos y deficiencias en las conexiones (porcentaje de sesiones, perdidas, transacciones incompletas…), etc., obteniendo una valiosa información que permitirá mejorar los contenidos y servicios incluidos en el Website.

De este modo, es posible realizar un seguimiento de las conexiones al Website para analizar las pautas de comportamiento general de los usuarios: ¿cuánto tiempo se dedican a consultar nuestras páginas Web?, ¿se registran y nos facilitan sus datos de contacto?, ¿repiten sus visitas con frecuencia?, ¿cuáles son los momentos del día o los días de la semana con mayor actividad?, ¿desde qué países o zonas geográficas se obtienen un mayor número de visitas?, ¿se acogen a una determinada promoción?, ¿compran on-line?, etc.

Asimismo, la identificación de los visitantes a un Website permite crear una valiosa base de datos de marketing, que puede ser utilizada para ofrecer productos y servicios a medida, personalizar el proceso de comunicación y desarrollar otras técnicas incluidas dentro del Marketing "One-to-One".

También es posible definir el perfil de cada uno de los visitantes a partir del seguimiento de sus pasos dentro de un Website en sus distintas conexiones: qué páginas visita, cuánto tiempo le dedica a cada una de las secciones, qué productos y servicios busca dentro del catálogo de la empresa, etc. Con toda esta información los responsables del Website podrán adaptar el contenido y la información en función de cada uno de los perfiles generados.

4. ANÁLISIS DE LOS REGISTROS DE ACTIVIDAD

Mediante el análisis de los "logs" (registros de actividad) del servidor Web es posible determinar, de forma no intrusiva (es decir, sin tener que preguntárselo directamente ni incomodar al usuario), la siguiente información:

- Fecha y hora en la que tiene lugar cada visita: permite obtener la distribución semanal y distribución horaria del tráfico de visitas al Website.
- Dirección IP de la máquina del visitante: permite averiguar el país, el dominio y la organización a la que pertenece.
- Dirección URL de la que procede ("referer log"): permite detectar si el usuario proviene de una página Web en la que la empresa ha insertado un banner u otro elemento publicitario, de una consulta por palabra clave en un buscador, de la página Web de un miembro de un programa de afiliación, etc.
- Idioma, tipo de navegador y tipo de sistema operativo utilizado en el equipo del usuario que visita el Website.
- Páginas Web solicitadas y tiempo dedicado a cada visita: facilita el análisis de qué contenidos son los más consultados y cuáles los menos interesantes dentro del Website, así como cuáles son los recorridos típicos por el Website, los puntos de entrada y los puntos de salida hacia otras direcciones de Internet.
- Códigos de respuesta, estado de la conexión y cantidad de información enviada por el servidor atendiendo a cada petición del cliente.

A partir de estos datos es posible obtener otra información de interés. Así, por ejemplo, el navegador Internet Explorer solicita el icono "favicon.ico" al servidor Web cada vez que el usuario decide registrar una página Web en su carpeta de Favoritos ("bookmarks"). Mediante el control de este tipo de peticiones, una organización podrá conocer qué visitantes de su Website han decidido registrar una o varias páginas en su carpeta de Favoritos.

También es posible realizar un seguimiento de la lectura y utilización de los mensajes de correo electrónico en formato HTML, ya que el servidor Web puede registrar cada una de las peticiones realizadas desde el equipo del usuario, cuando éste decide activar un determinado enlace incluido en el cuerpo del mensaje de correo electrónico. De este modo, se puede obtener un conocimiento mucho más detallado de los resultados de una campaña de comunicación a través del correo electrónico.

Una empresa podría generar distintas versiones de sus páginas Web, con diseños alternativos e incluyendo distintos tipos de contenidos, para poder hacer un seguimiento y evaluación en tiempo real de cuáles son las páginas Web

que tienen una mayor aceptación entre sus usuarios, seleccionando, en consecuencia, aquéllas que permitan mejorar la experiencia del usuario y la navegabilidad dentro del Website de la organización.

Disponemos en la actualidad de multitud de aplicaciones comerciales y gratuitas para el análisis del tráfico, como el software Webtrends Analyzer (www.webtrends.com), que ofrece datos y estadísticas sobre el número total de páginas vistas, el número total de visitas, el número de nuevos usuarios, los puntos de entrada en el Website, los principales recorridos por el Website, los tiempos medios de visita, los países y organizaciones de procedencia, etc.

Existen, no obstante, una serie de posibles factores de distorsión que afectan tanto al control de las visitas a un Website como a la descarga de elementos publicitarios on-line:

- La técnica de "caching", empleada en los proveedores de acceso a Internet y en los servidores "proxies" de las empresas: mediante esta técnica cada página Web que haya sido solicitada en una ocasión se guarda en el disco duro del servidor del proveedor o de la empresa, para acelerar posteriores accesos a esa misma página, ya que, de este modo, ésta no tendrá que ser descargada nuevamente desde el servidor Web original. Esta característica provoca que no se registren en dicho servidor Web las visitas de usuarios que acceden a las páginas que se encuentran almacenadas en la memoria "caché" de un "proxy".
- Utilización de agentes para buscar información: la página Web es leída y procesada por un programa informático (agente), que selecciona la información que pueda ser de interés para el usuario que representa.
- Navegación en modo texto: con esta opción del navegador no se visualizan los elementos gráficos ni las animaciones, por lo que también se eliminan los banners publicitarios de la página Web.
- Utilización de software inteligente para filtrar los banners publicitarios y otros contenidos no deseados por el usuario.
- Errores en la transmisión, que impidan una descarga completa de la página Web.
- Interrupción por parte del usuario de la descarga de los elementos de la página Web.

Para paliar en parte algunos de estos problemas y, en especial, la técnica de "caching", se pueden incluir marcadores HTML (un tipo de etiquetas construidas en HTML) en las páginas Web, que soliciten algún elemento del servidor Web (generalmente una imagen) en el momento en que éstas vayan a ser leídas en un navegador, aunque se encuentren guardadas en la caché de otro servidor o en el disco duro del ordenador del usuario. De este modo, cada lectura de una página Web podrá ser contabilizada en el servidor, siempre y cuando el usuario se encuentre en ese momento conectado a Internet.

5. IDENTIFICACIÓN DE VISITANTES

En la práctica se pueden utilizar distintas técnicas para identificar a los visitantes de un Website:

5.1. CONTROL DE LA PROCEDENCIA A PARTIR DE LA DIRECCIÓN IP

Es un procedimiento muy sencillo, ya que en cada conexión a un servidor Web el equipo cliente da a conocer su dirección IP. De este modo, registrando las direcciones IP de los visitantes es posible hacer un seguimiento de sus conexiones al servidor Web.

Sin embargo, no es una técnica demasiado fiable, debido a varios factores:

- Direcciones IP compartidas a través de un servidor "proxy": en muchas empresas la conexión a Internet se establece a través de un servidor "proxy", que actúa como intermediario entre las peticiones de la red interna e Internet, utilizando una única dirección IP que será compartida por todos los equipos conectados desde la red de la empresa.
- Direcciones IP dinámicas: La mayoría de los usuarios particulares que establecen su conexión a través de un proveedor de acceso no tienen una dirección IP fija, sino que ésta es asignada de forma dinámica en función de la disponibilidad de direcciones en el proveedor. Por este motivo, es muy probable que la dirección cambie de unas conexiones a otras. Sin embargo, con las conexiones a través de ADSL o cable-módem cada vez es más habitual emplear una dirección IP fija.

- Utilización de varios proveedores de acceso a Internet: con la proliferación de las ofertas de acceso gratuito a Internet, es bastante frecuente configurar varias cuentas con distintos proveedores y utilizar la que ofrezca un mejor rendimiento en función del tráfico en la red de cada proveedor. Además, un mismo usuario podría utilizar un proveedor para navegar desde el trabajo y otro distinto para hacerlo desde su propio domicilio.

5.2. UTILIZACIÓN DE 'COOKIES'

Las 'cookies' son ficheros guardados en el disco duro del visitante que registran datos sobre su navegación por las páginas de un servidor Web y lo identifican en posteriores conexiones.

Cada cookie es un pequeño fichero de texto (puede ocupar un tamaño máximo de 4 Kbytes) que no tiene capacidad de ejecución, por lo que no puede incluir código dañino en forma de virus que pueda representar algún riesgo para el equipo. Este fichero se guarda en el disco duro del ordenador del usuario. Para ello, el navegador Internet Explorer utiliza una carpeta denominada "Cookies" para cada usuario registrado en el sistema, mientras que el navegador Netscape los almacena en un único fichero denominado "cookies.txt".

Asimismo, cada cookie tiene una fecha de caducidad, de tal modo que será eliminado del disco duro una vez se haya alcanzado esa fecha. Pertenecen a un determinado servidor Web y no pueden ser leídos por un ordenador distinto del que lo envió al equipo del usuario (salvo fallos de seguridad en el navegador del usuario o ataques del tipo Cross-Site Scripting, XSS). Por otra parte, el usuario puede rechazarlos si así lo desea, estableciendo una determinada opción en la configuración del programa navegador.

Para su creación en el equipo del usuario, el servidor envía la información para crear cada cookie en una respuesta a una petición HTTP del equipo del usuario. A través de una cabecera especial en la respuesta (Set-Cookie), se definen campos como los siguientes:

- "Expires": fecha de caducidad de la cookie.
- "Domain": dominio asociado a la cookie.
- "Path": indica las páginas que deberían motivar el envío de la cookie.

El navegador enviará la cookie al servidor cuando el usuario realice nuevas conexiones a dicho servidor Web.

La técnica de las cookies fue desarrollada por la empresa Netscape para mejorar las capacidades de las aplicaciones cliente/servidor basadas en el Web (hay que tener en cuenta que el protocolo HTTP es un protocolo sin estado, que no recuerda anteriores fases de la conexión a un servidor Web).

Gracias a las cookies, el servidor Web puede recordar algunos datos sobre el usuario que le visita: sus preferencias para la visualización de las páginas en ese servidor (personalización de servicios y contenidos), cuál es su nombre de usuario y contraseña, etc.

Sin embargo, mediante las cookies no es posible recabar datos personales de los usuarios (nombre, apellidos, dirección de correo electrónico, etc.) y tampoco constituyen un método totalmente fiable para la identificación de usuarios: permiten identificar al ordenador que realiza la conexión, pero este equipo podría ser compartido por varias personas.

Por lo tanto, entre las principales aplicaciones de las 'cookies', podríamos citar:

- Control de los accesos a un Website: obtención de una información más exacta sobre el número de visitantes a cada página Web, distinguiendo entre el número de visitas y el número de impresiones.
- Personalización de la apariencia, servicios y contenidos de un Website.
- Implementación de carritos de la compra virtuales (shopping carts): las 'cookies' le permiten al servidor Web recordar los artículos que el cliente va seleccionando en una tienda virtual, para poder realizar el pedido de forma conjunta al final del recorrido por la tienda.
- Almacenamiento de información sobre el medio de pago empleado en las compras dentro de una tienda virtual: guardar en una 'cookie', por ejemplo, el número de la tarjeta de crédito del cliente para facilitar posteriores compras. No obstante, esta aplicación no es muy recomendable desde el punto de vista de seguridad, sobre todo si no se guarda la información encriptada.

También se han planteado varias objeciones a la utilización de 'cookies', sobre todo cuando no se informa al visitante de un Website de esta técnica:

- Posible violación de la intimidad del usuario: determinación de su ideología, aficiones o áreas de interés a través del seguimiento de sus hábitos de navegación en un Website.
- Almacenamiento de información sensible sobre el usuario en su disco duro, que podría plantear problemas de seguridad (algún intruso podría tener acceso al disco duro y obtener esa información).
- El usuario medio de Internet no es consciente de qué son las 'cookies' y cuál es su aplicación en cada caso.
- Mala utilización de la información por parte de la empresa, que podría generar una base de datos de perfiles con los datos recabados, para obtener un beneficio con su cesión a terceros. Esto está totalmente prohibido en los países de la Unión Europea, hasta el punto de que en España supondría una grave infracción de la Ley Orgánica de Protección de Datos (LOPD). Pero en otros países, como Estados Unidos, este tipo de prácticas empresariales no están penalizadas por la ley.

5.3. USUARIOS REGISTRADOS MEDIANTE UN NOMBRE Y UNA CONTRASEÑA

En este caso se solicita al usuario un registro previo para poder acceder a los servicios del Website, cubriendo para ello los distintos campos de un formulario de inscripción. Sin duda, ésta es la opción más fiable, ya que permite identificar a los visitantes aunque naveguen utilizando distintas cuentas de acceso o distintos equipos.

Sin embargo, es necesario ofrecer al usuario algún beneficio a cambio de sus datos y del tiempo que nos dedica: "vender" la posibilidad de personalizar el contenido del Website para enriquecer el proceso de comunicación; entregar un obsequio o la muestra de un producto; participar en un sorteo; etc.

En cuanto a las características del formulario de inscripción, se debería limitar la información solicitada para que se pueda completar rápidamente (en 2 ó 3 minutos), recabando información accesible y de respuesta inmediata, incluyendo, si es posible, preguntas que faciliten una segmentación posterior.

Por otra parte, se debe prestar especial atención al tratamiento de los datos capturados y la protección de la información personal, respetando las exigencias de la Ley Orgánica de Protección de Datos (LOPD).

6. TÉCNICAS DE WEB MINING

En lo que se refiere al análisis de los datos registrados en los "logs" de los servidores Web, algunos autores como Jackson (2002) han destacado la posibilidad de utilizar herramientas de Datamining.

De este modo, podemos utilizar el término de "Web Mining" para referirnos a la utilización de las herramientas y técnicas de minería de datos para descubrir y extraer información sobre la utilización de los servicios del Website de una organización. Otros autores también emplean el término "clickstream" para referirse al análisis de la utilización del Website.

Dentro del "Web Mining" podemos distinguir las siguientes áreas de trabajo:

- ❖ Minería de la Estructura del Website (Website Structure Mining), que trata de analizar cuál es la estructura real de un Website a través del estudio de los enlaces estáticos y dinámicos entre sus distintas páginas Web, contenidos y servicios.
- ❖ Minería del Contenido del Website (Web Content Mining), centrada en la recopilación de datos e identificación de patrones relativos al acceso a los contenidos del Website por parte de sus usuarios (minería de páginas Web), así como a las búsquedas que éstos realizan en Internet (minería de resultados de búsqueda).
- ❖ Minería de la Utilización del Website (Web Usage Mining), que parte de los registros de los servidores Web para analizar la navegación y las distintas transacciones realizadas por los usuarios, a fin de determinar sus patrones de uso: páginas más visitadas, recorridos habituales, lugar por donde comienza y por donde finaliza la visita, tiempo medio de visita, otros sitios remitentes, etc.

Podemos localizar distintas herramientas en Internet, algunas de ellas gratuitas, que facilitan

las técnicas de Web Mining, entre las que podríamos citar:

- Net Tracker (http://www.sane.com).
- WebTrends (http://www.webtrends.com).
- WebSideStory (http://www.websidestory.com).
- Blue Martini (http://www.bluemartini.com).
- Coremetrics (http://www.coremetrics.com).
- Elytics (http://www.elytics.com).
- Accrue (http://www.accrue.com).

Por lo tanto, mediante el análisis de los datos específicos del usuario (sobre todo cuando éste se ha registrado, indicando algunos datos de tipo sociodemográfico), de los datos sobre su entorno tecnológico (navegador, sistema operativo, resolución de la pantalla, software instalado, tipo de conexión, etc.) y de los datos acerca de la utilización y navegación dentro del Website, es posible implantar servicios y contenidos personalizados, como los sistemas para la recomendación de los productos que mejor se ajusten a las necesidades de cada usuario.

No obstante, pese al espectacular crecimiento del número de Websites disponibles en Internet, todavía un porcentaje muy bajo de éstos recurren al Web Mining para analizar su estructura, su contenido y su utilización, con el objetivo de poder mejorar el servicio ofrecido y la experiencia de sus usuarios.

RESEÑA CURRICULAR DEL AUTOR

Álvaro Gómez Vieites

Ingeniero de Telecomunicación por la Universidad de Vigo. Especialidades de Telemática y de Comunicaciones. Número uno de su promoción (1996) y Premio Extraordinario Fin de Carrera.

Ingeniero Técnico en Informática de Gestión" por la UNED (2004-2006). Premio al mejor expediente académico del curso 2004-2005 en la Escuela Técnica Superior de Ingeniería Informática de la UNED

"Executive MBA" y "Diploma in Business Administration" por la Escuela de Negocios Caixanova.

Ha sido Director de Sistemas de Información y Control de Gestión en la Escuela de Negocios Caixanova. Profesor colaborador de esta entidad desde 1996, responsable de los cursos y seminarios sobre Internet, Marketing Digital y Comercio Electrónico.

Socio-Director de la empresa SIMCe Consultores, integrada en el Grupo EOSA.

Autor de varios libros y numerosos artículos sobre el impacto de Internet y las TICs en la gestión empresarial.