Web Mining y personalización en tiempo real

Juan Bernabé Moreno
Universidad de Granada, Departamento de Ciencias de la Computación e Inteligencia Artificial

Abstract— **Este trabajo se centra en demostrar el papel de las técnicas de web mining en la personalización de documentos web. Tras repasar los tipos de web mining y los mecanismos de reunir y analizar los datos relacionados con la interacción de los usuarios con el sistema, se presenta un Framework para la implementación de técnicas de personalización con la peculiaridad de que estas personalizaciones deben satisfacer requerimientos de tiempo real.**

***Index Terms*— web mining, personalización, online marketing, tiempo real**

I. INTRODUCCIÓN

E-commerce representa hoy en día uno de mercados con mayor potencial y uno de los canales de transacciones económicas con mayor proyección. La variedad y cantidad de información y productos ofertados disponible en Internet deja patente la necesidad de las páginas web de dirigirse a usuarios con distintos intereses. Aunque está claro que la oferta y variedad de productos, servicios e información en Internet aumenta, todavía queda mucho terreno por delante para saber qué ofertas son hechas a qué usuarios con qué resultados.
El reto se puede ver como una moneda de dos caras:
- alcanzar según las ofertas de una página, los usuarios "apropiados"
- presentar a los usuarios las ofertas "apropiadas"
En ambos casos está claro qué usuarios vienen a la página y cómo se usa ésta –al contrario que en el comercio clásico, las páginas web ofrecen la posibilidad de analizar el uso de las ofertas y cómo hacerlas más atractivas o más específicas para los diferentes grupos de usuarios.
La información básica para el análisis de la interacción con las páginas web reside en los logs del servidor web, en los que se protocolan masivamente informaciones relevantes a este respecto. Dependiendo del tráfico y de la complejidad de la página web la información capturado en los logs diarios puede llegar a ocupar varios gigas.

Durante la última década se han desarrollado herramientas para evaluar la información de los logs, como FastStats Analyzer, WebTrends, WebSuxess, LogAnalyzer, etc. Estas herramientas deberían dar respuesta a las preguntas siguientes:

- Quién visita la página? Información demográfica, sobre el ISP, etc
- En qué zona horaria y temporada donde la página es accedida con más frecuencia
- Cuál es la distribución del tráfico en la página?
- Qué páginas son accedidas por qué usuarios (clientes, concurrencia, etc)?
- Qué navegador ha sido empleado para visitar la página?

Si uno quiere pasar de este simple análisis descriptivo del uso de la página, para por ejemplo poder analizar el comportamiento de los usuarios en la página, se requieren técnicas más detalladas que pertenecen al ámbito de la minería web (web mining).
Se distinguen tres componentes diferente dentro del Web Mining [2]

- Minería del contenido, que por ejemplo se centra en el descubrimiento de información relevante en el contenido de los documentos de la web (incluyendo información en diferentes formatos, accedida por diferentes protocolos, etc)
- Minería de la estructura web: se encarga de analizar cómo la página está linkada con otros recursos en Internet y se base en la topología de los hyperlinks.
- Minería de la utilización web: se encarga de analizar los datos que se graban en una sesión durante la interacción de un usuario con la página. Mientras las dos categorías anteriores utilizan datos primarios disponibles en la web, la minería de la utilización web se basa en los datos secundarios producidos en la interacción de los usuarios con la web, incluyendo los datos del servidor web, proxies, navegadores, perfiles de usuario, datos de registro, sesiones de usuario o transacciones, cookies, queries de usuario, Mouse clicks y scrolls, y cualquier otro tipo de información resultado de una interación.

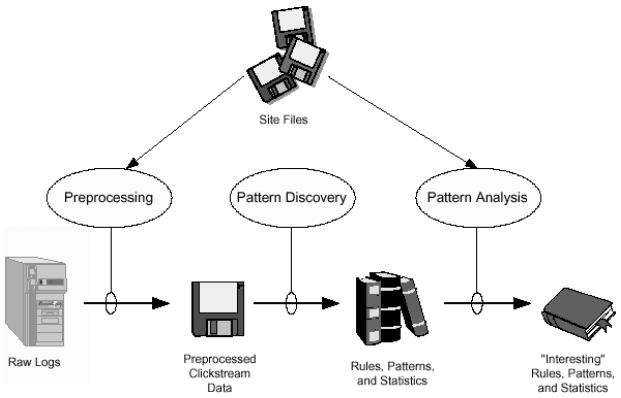

Figure 1 Procesos de la minería de uso web [3]

Según la Figura 1, la preparación de los datos es un paso crucial para todo el proceso, pero la mayor parte de la literatura sobre el tema se centra en el propio análisis de los datos

El trabajo presente analizaremos las posibilidades existentes para el análisis de los datos del Server que se centra en los datos derivados de la interacción de los usuarios con la página. En concreto, el objetivo será la generación de sistemas de reglas a partir del traceado de la actividad de los usuarios, a partir de las cuales las preferencias de cada grupo de usuarios pueden ser derivadas. Usando estas reglas como base para dirigirse a los clientes teniendo en cuenta la segmentación de usuarios en grupos se podrá conseguir el deseado nivel de personalización del sitio web para pasar de un modelo de márketing 1-to-n a uno 1-to-1.
Los posibles campos de aplicación son:
- optimización del diseno del contenido (CMS)
- aplicación del model Marketing-Controlling
- generación de perfiles de usuario y extracción de los click paths para analizar la conducta de los usuarios en la página
- derivación de Business rules para determinar la personalización más apropiada de la página

II. ANÁLISIS DE LOS LOGS

A. *Análisis de los logs files*

El uso de un web site queda protocolado en el web Server en forma de log files. Las entradas de los logfiles suelen contener la dirección IP del usuario, cuándo se ha producido el acceso, qué HTML comando se ha empleado (GET, POST, etc), la URL de la página solicitada, el protocolo (http, https, etc), el código de estado, el número de bytes transmitidos y en algunas ocasiones el ID del usuario [4]
No todas las entradas del log file son relevantes. Por ejemplo, todos los GET Request originados para devolver los elementos gráficos de una página suelen ser filtrados en un preproceso de los logs antes de emplear técnicas de web mining.

```
192.168.114.201, -, 03/20/01, 7:55:20, W3SVC2, SERVER, 172.21.13.45, 4502, 163, 3223, 200, 0, GET, /DeptLogo.gif, -,
```

Figure 2: Ejemplo de entrada en ficher de logs (IIS 6.0)

Table 10.6 Example of an IIS Log File Entry

Field	Appears As	Description
Client IP address	192.168.114.201	The IP address of the client.
User name	-	The user is anonymous.
Date	03/20/01	This log file entry was made on March 20, 2001.
Time	7:55:20	This log file entry was recorded at 7:55 A.M.
Service and instance	W3SVC2	This is a Web site, and the site instance is 2.
Server name	SERVER	The name of the server.
Server IP	172.21.13.45	The IP address of the server.
Time taken	4502	This action took 4,502 milliseconds.
Client bytes sent	163	The number of bytes sent from the client to the server.
Server bytes sent	3223	The number of bytes sent from the server to the client.
Service status code	200	The request was fulfilled successfully.
Windows status code	0	The request was fulfilled successfully.
Request type	GET	The user issued a **GET**, or download, command.
Target of operation	/DeptLogo.gif	The user wanted to download the DeptLogo.gif file.
Parameters	-	There were no parameters passed.

Figure 3: Información derivada de una entrada en el server log

Normalmente se encuentran en los log files la dirección IP de los usuarios que acceden como única medida de identificación de los usuarios. El problema es que asignar una IP a un usuario concreto no es trivial porque la IP interna del usuario es a menudo transformada en otra IP por los sistemas de red entre el ordenador que solicita la página, y el Server que alberga la página.

Esta situación hace necesario el análisis de la interacción de un usuario con la página durante el tiempo en el que una sesión es establecida. La sesión resume el tiempo de visita de un usuario a una página… lo que se podría identificar como el tiempo que pasa un cliente en una tienda. Durante la sesión, se transmiten click-streams, que identifican la duración de la visita, el grupo de usuarios al que pertenece el cliente, etc.

Existen varios métodos de identificación de un usuario con una determinada sesión:

1) Cookies

El propio servidor web o un servidor dedicado a este propósito pueden activar una cookie en el navegador del usuario. Esta cookie se envía junto con cada request al servidor web permitiendo así la localización de un determinado identificador de sesión. Existen dos variantes:
- Transient cookies: que no son almacenadas por el navegador y se pierden cuando el browser se cierra
- Persistent cookies: se almacenan en el equipo del usuario. Esto permite que el usuario pueda ser identificado en cada visita a la página (más allá de cada sesión particular)

2) URL Rewriting

Las cookies deben ser aceptadas por el usuario, y algunas veces no se llegan a aceptar o se desactivan, etc. En lugar de las cookies, se puede reconocer el identificador de cada sesión añadiéndolo como parámetro a cada URL solicitada. De esta manera se puede tracear la sesión que corresponde a cada request.

3) Java Applets

El empleo de determinados Applets permiten no sólo transmitir el identificador de sesión con cada request, sino también información adicional sobre el usuario. Como las cookies, requieren que el usuario especifícamente acepte el hecho de que el Applet envíe esta información al Server.

En general en los logs del servidor no se encuentra información alguna sobre la interacción del usuario con el applet… a lo sumo, que el applet ha sido invocado. Los applets tienen que ser extendidos para protocolar la interacción del usuario, información que luego ha de ser comparada con la que se encuentra en los logs de servidor.

4) Identificación de sesión en el servidor de aplicaciones

Los portales, sites y tiendas online más complejos requieren al mismo tiempo aplicaciones más complejas que corren en un servidor de aplicaciones que se comunica con un servidor web. Para reproducir la conducta de los usuarios, se requiere invocar en ciertos puntos de la aplicación un Tracking Server. Se trata principalmente de *Business Events* o eventos que deciden el flujo de la aplicación.

5) Reverse Proxy Server

La función de un reverse Proxy Server consiste en redistribuir las requests entre diferentes servidores web internos. Para el cliente, sólo existe un servidor y no se percata de que diferentes partes de la aplicación o diferentes aplicaciones se encuentran en diferentes servidores. Hay que resaltar, que para los servidores situados detrás del Reverse Proxy es muy difícil averiguar qué usuario envió el request, porque a éste sólo le llegan requests del Proxy Server.

Cabe resaltar que independientemente del mecanismo elegido para registrar la interacción de los usuarios con la página, el empleo de técnicas de caching y de elementos dinámicos en las páginas web dificultan el análisis completo de la conducta de los usuarios.

Caching consiste en mantener versiones de la página en diferentes servidores de caching para reducir el tiempo de respuesta. La consecuencia inmediata es que el operador del site pierda la posibilidad de registrar ciertas interacciones de usuarios con las páginas, al ser servidas desde un caching Server. Se puede indicar en el http header la directiva NO-CACHE para evitar este escenario, pero es recomendable analizar el equilibrio información registrada vs. Tiempo de respuesta para cada página concreta.

En el caso de las páginas dinámicas, especialmente en portales donde el usuario dispone de un alto grado de configuración del layout, se hace difícil el registro de los requests individuales y la asociación con un identificador de sesión concreto. En estos casos se requiere un mecanismo de logs a nivel de aplicación adicionalmente a los logs de servidor, pero integrar ambos no es una tare trivial

III. Personalización

Las disciplinas E-Commerce y Marketing otorgan cada vez más importancia al análisis de las búsquedas y las decisiones de clientes potenciales. El objetivo es dirigirse a los clientes de manera óptima, ya sea reflejada en la manera de colocar los productos en una estantería de una tienda, o en los videos que se proyectan en un spot de publicidad. Internet no es una excepción si se contempla como un canal comercial.

Personalización se puede definir como mostrar los contenidos apropiados para un determinado usuario, según la información disponible de éste usuario, por ejemplo el historial de compras, la interacción del usuario con la página o información adicional disponible sobre esa persona [6]

A. Personalización a través del usuario

Diferentes aplicaciones de Internet ofrecen la posibilidad para el usuario de definir el diseño y/o las ofertas del site según sus necesidades. Esta forma de personalización se denomina Creación de Perfiles o Individualización[7]

B. Personalización basada en sesiones

Durante una sesión se puede adaptar la oferta en información y en productos que se hace al usuario. Por ejemplo YouTube ofrece videos que pueden interesar al usuario basándose en los videos que el usuario ya ha visto. Amazón hace lo mismo según el historial de compras. Es importante destacar que las reglas para determinar qué contenido y qué productos se ofrecen a qué usuarios se defines offline[7]

C. Personalización más allá de las sesiones

Mientras sea posible identificar al usuario por ejemplo a través de un login y un password, se puede personalizar la experiencia del usuario con la página. Por ejemplo en e-Banking, dependiendo de los datos disponibles de un cliente – estado civil, disponibilidad de un automóvil, edad, etc- se pueden personalizar las ofertas de créditos, seguros, fondos de inversión, etc.

Hay que reseñar que aspectos relacionados con la protección de datos deben tenerse en cuenta, como por ejemplo el principio de anonimicidad que indica que los datos de un usario deben almacenarse separados de tal forma que no identifican a la persona física [7]

IV. Web Mining y filtrado colaborativo

Personalización implica conocer y anticiparse a las necesidades del usuario. Esto implica pronosticar sus patrones de navegación, sus solicitudes de información y su intención de compra de productos. Cada transacción o navegación se puede interpretar como un evento disponible en los logs de aplicación cuyo resultado se puede consultar de manera estructurada en una base de datos, y todo este cúmulo de informaciones puede ayudar a ajustar la página usando ofertas personalizadas, modificaciones del diseño de la página o

banners dirigidos a diferentes grupos de usuarios. El empleo de técnicas de minería de datos se hace imprescindible según la complejidad de la personalización aumenta y se hace insostenible el modelo del hard-codeo (redes neuronales, algoritmos basados en árboles de decisión o técnicas de análisis multivariable para extraer perfiles de usuario de todos los datos productos de la interacción de los usuarios con la página). Sistemas de reglas permiten describir los perfiles y determinar qué actividades se ofrecen a qué grupo de usuarios.

De una manera muy simplificada se puede describir la minería web como la aplicación de algoritmos de la minería de datos sobre los logs disponibles –de aplicación y de servidor-, que representan las actividades y transacciones de los usuarios.[8] Los logs se suelen almacenar de manera estructurada en un Web Data Mart y se presentan de la manera más apropiada según los algoritmos de minería de datos seleccionados. Algunas técnicas conocidas se presentan a continuación:

A. *Algoritmos de Análisis de secuencias de los Click Streams*

Utilizando algoritmos de análisis de secuencia se pueden extraer patrones de navegación. De esta manera se pueden por ejemplo insertar banners en los trayectos de navegación más populares u optimizar la ergonomía de la página[9]

B. *Clustering para identificar los perfiles de usuario*

La existencia de diferentes perfiles es la base de toda personalización. Los perfiles o clusters se describen mediante reglas que se almacenan en el servidor de aplicaciones. Tan pronto como un usuario es reconocido como perteneciente a un perfil, recibe contenido específico para ese perfil.

C. *Árboles de decisión*

Para cubrir dependencias más complejas dentro de un site, se crean modelos de clasificación y pronóstico basados en árboles de decisión, que no hacen otra cosa que representar un conjunto de registros en forma de árbol que se puede describir con reglas del tipo "SI … ENTONCES …" Por ejemplo se pueden obtener informaciones sobre la probabilidad de cerrar una compra de un producto analizando diferentes parámetros como duración de la sesión, patrones de navegación, etc. [9]

Los sistemas de recomendación pertenecen a las técnicas adaptativas o de asistencia al visitante de la página para encontrar de una manera más fácil y rápida lo que busca. Estos sistemas explotan los datos procedentes de la interacción de otros usuarios pertenecientes al mismo perfil con la página para hacer recomendaciones a un determinado usuario.

Estos sistemas requieren un periodo inicial en el que se acumulan datos y se aplican técnicas de clustering para perfilar los grupos de usuario.

V. FRAMEWORK PARA LA PERSONALIZACIÓN[1]

La concurrencia feroz y la velocidad a la que cambio de la nueva economía cristalizan en nuevos requisitos para los sistemas IT, los procesos que los gobiernan y las organizaciones que los implementan. La velocidad de reacción es decisiva sobre los competidores y ello implica la disponibilidad de información actual y de alta calidad para suportar los procesos de decisión.

Las herramientas para evaluar la interacción de los usuarios con el site se hacen indispensables para analizar posibles mejoras y evaluar la reacción de los usuarios a los cambios introducidos. El procedimiento general persigue protocolar la conducta de los usuarios a partir de los logs de aplicación y los logs de servidor. La calidad de los logs de aplicación y la disponibilidad en tiempo real de esta información en un sistema de gestión de base de datos permite generar informes y aplicar técnicas de minería de datos. Este procedimiento constituye un Framework que permite determinar la personalización de la página tras aplicar minería de datos.

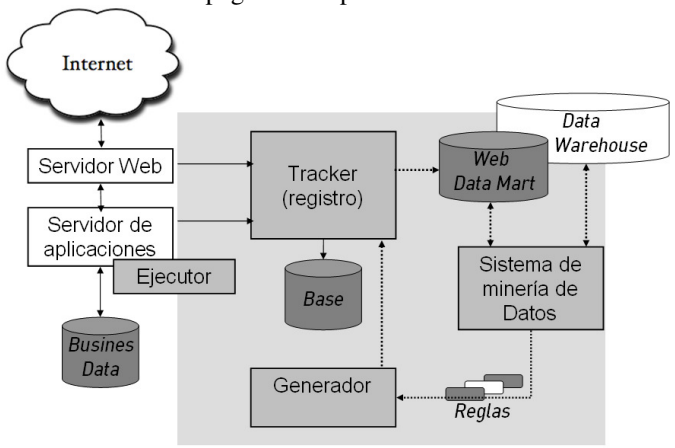

Figure 4 Arquitectura de un sistema de personalización

A. *Definición de objetivos*

Es crucial proceder orientados a un objetivo en el caso de un sistema de personalización. Se recomienda analizar antes de emprender ninguna personalización, lo que se quiere conseguir. Además es recomendable proceder de un modo gradual, por ejemplo personalizando ciertas partes de la aplicación. La ventaja de esta modularidad es una más rápida implementación de la personalización.

Una vez decidido que se quiere personalizar, el siguiente pase es determinar qué interacciones del usuario serán objeto de la personalización. La definición de eventos y su descripción detallada determinan según una determinada conducta del usuario qué actividad en forma de un banner o una determinada paleta de productos a ofertar deben aparecer en la página. Esto requiere registrar estos eventos durante una determinada sesión usando un Tracker. Los eventos pueden ser una determinada acción del usuario en la página o eventos con un transfondo de negocio, como la selección de determinados filtros de búsqueda de un producto.

El modelo presentado en la Figura 4 posibilita la definición de pasos dirigidos al ciclo de personalización, que comienza con el registro de información del usuario, sigue con la agregación de todos los datos y con el web mining, y finaliza

con la generación y aplicación de reglas para personalizar el contenido y los productos del site.

B. Tracking en tiempo real y construcción del Web Data Mart

Una vez definidos los eventos y el propósito de la personalización, la información necesaria para la página web queda determinada y por consiguiente la información que debe ser registrada por el Tracker. Los logs procedentes del Tracker se almacenan en el Web Data Mart y sirven de base para la minería de datos.

Para el Trucking se requiere que los datos de los eventos que se han definido para el tracking se transfieran al Tracker. Estoy puede tener un impacto en la velocidad de repuesta de la página. Normalmente se opta por una solución cliente servidor (javascript Trucking pixel) que mediante requests asíncronos no penalice la interacción con la página principal. El Tracker tiene la misión de agregar los logs de diferentes sistemas, como los del Server o los de aplicación, pero preservando la pertenencia a una determinada sesión pero medio del identificador de sesión.

La base de datos Web Data Mart contiene información acerca del número de clicks por sesión, duración de la sesión número de transacciones por sesión en una tabla central e información accesoria cubriendo por ejemplo datos relacionados con la fecha, la hora, el usuario logado, etc

Se pueden aplicar técnicas OLAP para analizar la información en todas sus dimensiones, que junto con los algoritmos de minería de datos establecen las bases para el control de las actividades online sobre la página.

Un aspecto muy importante es la necesidad de generar informes a partir de la información disponible en tiempo real.

VI. El bucle de la Personalización

La personalización se lleva a cabo por medio de dos bucles de información, el analítico y el de tiempo real.

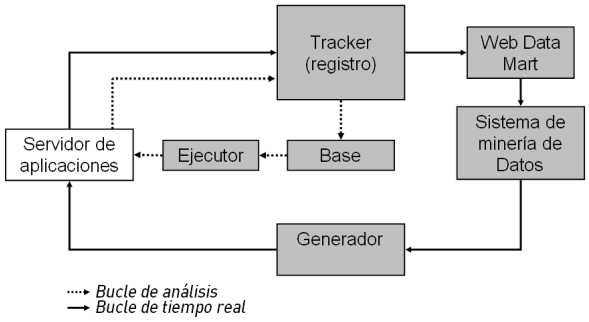

Figure 5 Bucles de la personalización

El bucle analítico se encarga de la preparación y el análisis de los logs disponibles cuyo resultado es una colección de reglas que describen los perfiles de los usuarios que visitan la página.

Estas reglas se pueden interpretar en la aplicación propiamente dicha para personalizarla. Así por ejemplo pueden ofrecerse determinados productos a determinados groupos de usuarios, o el layout puede ajustarse a las necesidades de cada grupo. Las reglas se basan en datos que se han agregado y transformado que no se disponen directamente en la aplicación. De este modo, intentar generar y aplicar reglas directamente a partir de la información de los logs disponible en la aplicación en tiempo real no es factible.

El bucle de tiempo real se encarga de la personalización en tiempo real. El generador facilita los datos resultado de aplicar las reglas del bucle analítico a los datos resultantes de la minería de datos. El ejecutor es el componente que se encarga de aplicar las reglas definidas en XML y mantenidas en la Base. Tras clasificar el usuario conectado, facilita el contenido personalizado. Para cerrar el círculo diremos, que los datos del usuario conectado son transmitidos nuevamente al Tracker, empezando de nuevo el proceso.

Conclusiones y Posibles líneas de investigación

En este trabajo se ha introducido la minería web, la personalización y se ha demostrado cómo ambas disciplinas están estrechamente relacionadas.

Se han presentado las distintas formas de obtener información de la actividad de los usuarios en una determinada página web o en un site completo, y se han discutido los distintos procedimientos de identificar el usuario que interactúa con el sistema.

Finalmente se ha movido el foco a la personalización en tiempo real y se ha presentado un Framework para implementarla de manera óptima.

Durante toda la exposición ha quedado patente el papel incipiente de las técnicas de personalización para el comercio electrónico y el marketing online.

Desde mi punto de vista, las perspectivas de la personalización se centran en el salto cualitativo que representa la web semántica tanto para la organización de la información registrada de los usuarios y las capacidades adicionales de razonamiento automático, como en las posibilidades de intercambio de datos de usuario entre aplicaciones traspasando la frontera de los dominios

References

[1] Meyer, M, Weingärtner, S, Jahke T., Lieven O. Web Mining und Personalisierung in Echtzeit. Heft 5 / 2001

[2] Kosala, R.; Blockeel, H.: Web Mining Research: A Survey. In: SIGKDD Explorations, Vol. 2, No. 1/2000, S. 1-15.

[3] Srivastava, J.; Cooley, R.; Deshpande, M.; Tan, P.-N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: SIGKDD Explorations, Vol. 1, No. 2/2000, S. 12-23

[4] Zaïane, O.R.; Xin, M.; Han, J.: Discovering Web Access Patters and Trends by Applying OLAP and Data Mining Technology on Web Logs. In: Proc ADL'98 (Advances in Digital Libraries), Santa Barbara, April 1998.

[5] Cooley, R.; Mobasher, B.; Srivastan, J.: Data Preparation for Mining World Wide Web Browsing Patterns. In: Knowledge and Information Systems, Vol. 1, 1/1999, S. 5-32.

[6] Röder, H.: Electronic Commerce und One to One-Marketing. In: Bliemel, F.; Fassott, G.; Theobald, A. (Hrsg.): Electronic Commerce. Gabler, Wiesbaden 1999, S. 213-224

[7] IBM High-Volume Website-Team: Personalisierung von Websites. http://www-106.ibm.com/developerworks/edeu/library /personalization.htm

[8] Berry, M.J.A.; Linoff, G.S.: Mastering Data Mining. Wiley, New York 2000.

[9] Weingärtner, S.: Web Mining – Ein Erfahrungsbericht. In: Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D. (Hrsg.): Handbuch Data Mining Marketing, Wiesbaden 2001, S. 889-903.