Sistema de Desarrollo de Estrategias de Marketing e Inteligencia de Negocios Usando Web Mining

Patricio Alcivar [1], Fanny Idrovo [2], Víctor Macas [3], Fabricio Echeverría[4].

[1] Ingeniera en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: palcivar@.espol.edu.ec
[2] Ingeniera en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: fidrovo@fiec.espol.edu.ec
[3] Ingeniero en Computación Especialización Sistemas Tecnológicos, 2006; e-mail: vmacas@fiec.espol.edu.ec
[4] Director de Tópico, Ingeniero en Computación, Escuela Superior Politécnica del Litoral, 1998, Profesor de ESPOL; e-mail: pechever@uniplex.com.ec

Resumen

Cuando los visitantes interactúan con su sitio, proveen información acerca de ellos y de como responden a su contenido: que enlaces visitaron, donde gastaron más su tiempo, y cuando navegaron. Algunos visitantes pueden incluso dar información de su estilo de vida o proveer nombres y direcciones, productos de competitividad o complementarios. Toda esta información es usualmente almacenada en una base de datos. Como resultado, se tiene mucha información de la Web, visitantes y contenido; pero probablemente no se esta haciendo el mejor uso de esta información. En este artículo describimos el uso de actividades de minería en la Web, que apuntan a la extracción de modelos del comportamiento navegacional de los usuarios de un sitio Web. Los modelos son inferidos del Log de un servidor Web por medio de datos y técnicas de Web Mining. La extracción de conocimiento es realizada con el propósito de ofrecer una vista personalizada y proactiva de los servicios Web para el usuario. Primero describimos el preprocesamiento, pasos en el Log necesarios para limpiar, seleccionar y preparar datos para la extracción de conocimiento. Luego, mostramos y explicamos las tres principales técnicas de Web Mining que usaremos: Reglas de Asociación , Clustering y Patrones Secuenciales.

Palabras Claves: *Web Mining ,extracción de conocimiento, log del servidor, reglas de asociación, clustering, patrones secuenciales.*

Abstract

When the visitors interact with your site, they provide information about themselves and how they respond to your content: which links visitors click, where they spend most of their time, and when they browse. Some visitors may even fill out a lifestyle survey or provide names and addresses, and features of competitive or complementary products. All this information is often stored in a database. As a result, you have a lot of information on your Web visitors and content, but you probably aren't making the best use of that information [1]. *In this article we describe the Web usage mining activities, that aims at extracting models of the navigational behaviour of a Web site users. The models are inferred from the access logs of a Web server by means of data and Web Mining techniques. The extracted knowledge is deployed to the purpose of offering a personalized and proactive view of the Web services to users. We first describe the preprocessing steps on access logs necessary to clean, select and prepare data for knowledge extraction. Then, we show and explain three main Web Mining techniques that we use: Association Rules, Clustering and Sequential Patterns.*

1. Introducción

Nuestro objetivo es desarrollar una aplicación que permita analizar archivos log generados por un servidor Web para encontrar patrones de navegación de usuario. Web Mining es simplemente aprovechar las técnicas de data mining para obtener conocimiento de la información disponible en Internet. Web Mining es una aplicación particular del data mining que busca patrones en hipertextos ampliada con la funcionalidad de considerar la estructura de los sitios Web. El análisis de estos patrones puede servir para personalizar el contenido del Web site, potenciar el mercado a través de una toma de decisiones eficiente en base a la salida que genere nuestro sistema, mejorar la estructura de la Web y, en definitiva, proporcionar un mejor servicio a los usuarios del sitio Web. Es importante señalar que muy a menudo los expertos que usarán los resultados del Web Mining son diferentes de los responsables que diseñan y construyen los modelos de Web Mining. Por ello se debe poner especial énfasis en la construcción de un sistema que proporcione unos resultados que se deja en manos de otros para su correcto análisis. En otras palabras, la salida que proporcione nuestro sistema debe ser fácilmente comprensible, portable y además, debe incluir la máxima información posible aprendida.

2. Web Mining

Una nueva corriente de investigación y desarrollo que ya ha comenzado a brindar algunos frutos y que seguramente afectará el diseño y la estructura tanto de los servidores como de los browsers de la Web.

Existen diferentes enfoques, en algunos casos se deben adaptar las técnicas conocidas de data mining para utilizarlas en este entorno, aunque en otros deben adaptarse los datos para que puedan ser utilizados.

Dentro del mundo de Web mining se pueden encontrar dos ramas de desarrollo bien diferenciadas, la que trabaja a nivel cliente y la que trabaja a nivel servidor de Web [2].

La primera se basa en aplicar mining sobre documentos obtenidos la red. Esto permite mejorar la búsqueda de información, generar perfiles de usuarios adecuados a sus necesidades y organizar bookmarks entre otras cosas.

La segunda aplica mining sobre los datos que dejan, en distintos tipos de logs, los Servidores de Web. Analizar esta información puede ayudar principalmente a empresas que basan su negocio en Internet determinando los tipos de clientes que ingresan, diseñar estrategias de marketing sobre productos y servicios, evaluar la efectividad de las campañas promocionales, mejorar tiempos de acceso y buscar la mejor estructura para el site.

Cada uno de estos enfoques presenta ventajas y desventajas, pero el enfoque en que basamos nuestra investigación es a nivel Servidor de Web que es el que analizaremos a continuación.

2.1 Web Mining en Servidores de Web

Actualmente los servidores Web generan un gran volumen de datos proveniente del registro de las acciones que estos realizan. Cada requerimiento de los clientes (browsers, agentes, entre otros) queda registrado en los logs que se generan constantemente.

Este gran volumen de datos contiene valiosa información que no es visible de forma evidente, y que hasta hace poco era utilizada mínimamente para obtener algún tipo de estadísticas, analizar accesos inválidos o problemas que se produjeran en el servidor [3].

Existen numerosas herramientas que generan reportes estadísticos y gráficos sobre el uso del servidor, de las cuales podemos destacar algunos productos conocidos como: Webtrends, Getstats, Analog, Microsoft Intersé Market Focus [2].

Ninguna de estas herramientas realiza data mining de los datos. No se aplica ni clusterización, ni reglas de asociaciones, aún menos sequential patterns. Con una herramienta de data mining se podría descubrir en el servidor: en forma general los clientes que realizan compras on-line habían consultado ciertas páginas los días anteriores. Con una herramienta estadística se podrían obtener totales por dominio, cantidad de requerimientos por recurso, entre otros.

En este momento, debido a la gran cantidad de negocios que se manejan por Internet, la gran competencia y la creciente necesidad de mejorar los servicios, el análisis de los datos que se obtienen para convertirlos en información útil se torna imprescindible para poder sobrevivir en este ambiente competitivo. Es necesario conocer el comportamiento de los usuarios (potenciales clientes) y brindarles un acceso más fácil y un mejor servicio así como también saber hacia quien orientar las campañas promocionales.

He aquí la importancia del Web mining en la toma eficiente e inteligente de decisiones en los negocios a través de Internet.

Además aprovechar esta información puede ser muy útil para mejorar el performance de los servidores.

Los datos almacenados en los logs siguen un formato standard diseñado por CERN y NCSA. Una entrada en el log siguiendo este formato contiene entre otras cosas, lo siguiente: dirección IP del cliente, identificación del usuario, fecha y hora de acceso, requerimiento, URL de la página accedida, el protocolo utilizado para la transmisión de los datos, un código de error, agente que realizó el requerimiento, y el número de bytes transmitidos. Esto es almacenado en un archivo de texto separando cada campo por comas (",") y cada acceso es un renglón distinto

Peo-il1-21.ix.netcom.com - - [24/Feb/1997:00:00:21 +0000] "GET /images/nudge.gif HTTP/1.0" 200 37 "http://www.internet.ibm.com/" "Mozilla/2.0 (compatible; MSIE 3.01; Windows NT)"
Slip166-72-149-200.wv.us.ibm.net - - [24/Feb/1997:00:00:21 +0000] "GET / HTTP/1.0" 200 9185 "http://www.ibm.com/Products/" "Mozilla/2.0 (Win95; I)"
ss5-08.inre.asu.edu - - [24/Feb/1997:00:00:21 +0000] "GET /commercepoint/html3/purchasing/3_a.html HTTP/1.0" 200 6277 "http://www.internet.ibm.com/commercepoint/html3/purchasing/3.html" "Mozilla/3.0 (Win95; I)"

Tabla 1. Ejemplo de entradas del log de NCSA

Estos datos contenidos en los logs son insuficientes para analizarlos directamente. Sin embargo utilizando una buena técnica de data mining se puede obtener información interesante.

Como fue mencionado anteriormente, se genera una entrada en el log por cada requerimiento de un recurso realizado por un usuario. Aunque esto puede reflejar la actividad del servidor no refleja el verdadero comportamiento de los clientes, debido a que las vueltas atrás y a que los requerimientos que se encuentran cacheados por el browser del cliente o por un proxy no son registrados en el server. Tampoco son registradas las funciones que el usuario realiza dentro de una página como por ejemplo el scroll-up y scroll-down. Este déficit en la información puede generar conclusiones erróneas al estudiar la mejor estructura para un site. Por ello dicha información debería ser registrada en logs generados por los browsers o por una applet Java.

A este déficit de información se suma que el identificador de usuario no siempre está disponible en el log. Debido al uso de proxy servers por parte de los proveedores del servicio de Internet y de firewalls por parte de las corporaciones comerciales, la verdadera dirección IP del cliente no se encuentra disponible para el servidor de Web. En vez de tener varias direcciones IP distintas para varios clientes distintos, la misma dirección del proxy server o firewall es guardada en el log representando los requerimientos de diferentes usuarios que llegan al servidor desde el mismo proxy server o firewall. Esto genera cierta ambigüedad en los datos del log. Para solucionar este problema, generalmente se requiere que los usuarios completen un formulario de registración, se implementa algún tipo de "log-in" o se utilizan "cookies" entre el servidor y el browser del cliente. De esta manera, el servidor puede identificar distintos requerimientos realizados por los usuarios, pero se "viola" la privacidad de los mismos ya que generalmente ellos desean permanecer anónimos lo más que se pueda. Por este motivo, los servidores no solicitan registros ni utilizan cookies; por lo tanto el análisis para identificar el comportamiento de los usuarios debe basarse sólo en las entradas del log.

2.1.1 El Proceso de Knowledge Discovery (KDD) Antes de aplicar cualquier técnica de data mining es necesario realizar una transformación de los datos para que éstos puedan ser operados eficientemente. A este proceso se lo conoce como el proceso de Knowledge Discovery [5]. En el marco de dicho proceso se filtrarán datos que no interesan y en general se transformará el log en una estructura más manipulable (por ejemplo una base de datos relacional). Es necesario el conocimiento de la estructura del Web server para poder determinar a partir de los accesos cual es la acción que quiere realizar el usuario.

Para poder aplicar las técnicas de data mining sobre los datos del log del servidor es necesario, además de aplicar las transformaciones en los datos típicas del proceso de KDD, realizar una adaptación en la definición de las transacciones y los ítems que las componen para los distintos algoritmos. Esto se debe a que en este caso no se tiene la noción de transacción como en una base de datos transaccional en donde existe un "identificador de transacción". Aquí para poder delimitar una transacción se debe utilizar una combinación entre el identificador del usuario que interactúa con el servidor y un período máximo de tiempo aceptado entre accesos. Si un usuario accede a una página del servidor a las 9:00 horas, y hasta las 9:15 horas navega dentro del site; y luego vuelve a acceder por la tarde, esto es considerado como dos transacciones distintas(Sesionización).

2.1.2 Técnicas de Data Mining En el Web mining en los servidores de Web es posible aplicar cualquiera de las técnicas de data mining conocidas: [4] Reglas de Asociación, Clusterización y Secuencia de Patrones; cada una de las cuales serán analizadas a continuación.

Reglas de Asociación

El descubrimiento de reglas de asociación [6] es generalmente aplicado a Bases de Datos transaccionales, donde cada transacción consiste en un conjunto de ítems. En este modelo, el problema consiste en descubrir todas las asociaciones y correlaciones de ítems de datos donde la presencia de un conjunto de ítems en una transacción implica (con un grado de confianza) la presencia de otros ítems.

En el contexto de Web mining este problema tiende a descubrir la correlación entre los accesos de los clientes a varios archivos disponibles en el servidor. Cada transacción está compuesta por un conjunto de URL accedidas por el cliente en una visita al servidor.

Utilizando association rules, se puede descubrir, por ejemplo, lo siguiente:

60% de los clientes que acceden a la página con URL /company/products/, también acceden a la página /company/products/product1.html.

Esta técnica, además, considera el soporte para las reglas encontradas. El soporte es una medida basada

en el número de ocurrencias de los ítems dentro del log de transacciones.

En Web mining existen otros factores que pueden ayudar a podar el espacio de búsqueda de las reglas. En general, los sites están organizados jerárquicamente y la estructura de esta jerarquía es conocida con anticipación. Por ejemplo, si el soporte de /company/products/ es bajo, se puede inferir que la búsqueda de reglas de asociación en las páginas /company/products/product1.html y /company/products/product2.html no van a tener el soporte necesario.

Además las reglas de asociación pueden ayudar a mejorar la organización de la estructura del site. Si descubrimos que el 80% de los clientes que acceden a /company/products y /company/products/file1.html también acceden a /company/products/file2.html, parece indicar que alguna información de file1.html lleva a los clientes a acceder a file2.html. Esta correlación podría sugerir que ésta información debería ser movida a /company/products para aumentar el número de usuarios que lo vean.

Clusterización

Las técnicas de clasificación permiten desarrollar un perfil para los ítems pertenecientes a un grupo particular de acuerdo con sus atributos comunes. Este perfil luego puede ser utilizado para clasificar nuevos ítems que se agreguen en la base de datos.

En el contexto de Web mining, las técnicas de clasificación permiten desarrollar un perfil para clientes que acceden a páginas o archivos particulares, basado en información demográfica disponible de los mismos. Esta información puede ser obtenida analizando los requerimientos de los clientes y la información transmitida de los browsers incluyendo el URL.

Utilizando técnicas de clasificación, se puede obtener lo siguiente:

50% de los clientes que emiten una orden on-line en /company/products/product2.html, están entre 20 y 25 años y viven en la costa oeste.

La información acerca de los clientes puede ser obtenida del browser del cliente automáticamente por el servidor; esto incluye los accesos históricos a páginas, el archivo de cookies, etc. Otra manera de obtener información es por medio de los registros y los formularios on-line.

La agrupación automática de clientes o datos con características similares sin tener una clasificación predefinida es llamada clustering.

Patrones Secuenciales

En general, se tienen disponibles los datos en un período de tiempo y se cuenta con la fecha en que se realizó la transacción; la técnica de patrones secuenciales se basa en descubrir patrones en los cuales la presencia de un conjunto de ítems es seguido por otro ítem en orden temporal.

En el log de transacciones de los servidores de Web, se guarda la fecha y hora en la que un determinado usuario realizó los requerimientos. Analizando estos datos, se puede determinar el comportamiento de los usuarios con respecto al tiempo.

El descubrimiento de patrones secuenciales en el log puede ser utilizado para predecir las futuras visitas y así poder organizar mejor los accesos y publicidades para determinados períodos. Por ejemplo, utilizando está técnica se podría descubrir que los días laborables entre las 9 y las 12 horas muchas de las personas que accedieron al servidor lo hicieron para ver las ofertas y en los siguientes días la mayoría compró productos. Entonces por la mañana debería facilitarse el acceso a las ofertas y brindar la publicidad más llamativa posible.

También puede ser utilizado para descubrir tendencias, comportamiento de usuarios, secuencias de eventos, etc. Esta información puede ser aprovechada tanto en el aspecto comercial (pensar una campaña de marketing) como en el aspecto técnico (mejorar los tiempos de acceso).

2.2- Beneficios y Problemas

En general aplicar técnicas de data mining sobre el log de los servidores puede brindar las siguientes ventajas:

- Mejorar la performance del servidor.
- Mejorar la navegabilidad del site
- Mejorar el diseño de las aplicaciones del Web
- Descubrir potenciales clientes de comercio electrónico
- Identificar lugares y horarios principales para colocar publicidades

Actualmente existen varios problemas que afectan a la exactitud de los resultados obtenidos al realizar el análisis. Entre ellos podemos destacar:

- Imposibilidad de registrar los accesos a páginas cacheadas o descubrir fehacientemente el uso del backtracking u otras funcionen del browser
- Dificultades en delimitar transacciones o sesiones del usuario
- Datos ambiguos en el log debido a cambios de identidad realizados por proxys y firewalls
- Estructura de los logs no adecuada para aplicar las técnicas de data mining

Como se puede ver, actualmente los logs no almacenan toda la información necesaria para hacer un buen análisis. Debido a que hace poco se comenzó con la aplicación de estas técnicas de data mining, logrando importantes avances, es de esperar que muy pronto se produzcan mejoras como la mayor cooperación entre browsers y servidores y posiblemente la adecuación de la estructura de los logs

para que éstos puedan ser analizados más eficientemente. Quizás, en el futuro, se puede lograr la aplicación de data mining on-line para adaptar rápidamente la estructura y la imagen de los servidores de acuerdo a las necesidades del momento.

Mientras tanto se debe encontrar la mejor manera de aprovechar los datos insuficientes y ambiguos con que se cuenta. Para ello es muy importante realizar un buen proceso de KDD aprovechando el conocimiento que se tenga sobre el dominio de la aplicación.

3.Conclusiones

En general, todos los algoritmos que se implementarán enfocan el análisis sobre secuencias de tiempo ya que los eventos que son almacenados están muy relacionados con el tiempo en que se producen.

El descubrimiento de estas reglas en el ámbito del comercio electrónico pueden ayudar en el desarrollo de las estrategias de marketing.

La utilización de la técnica de clusterización sobre el log del Web Server, puede ser utilizado para estrategias de marketing dirigido según las clases obtenidas. Por ejemplo si se reconoce un grupo de potenciales clientes se les podría enviar las ofertas por correo sólo a ellos.

Referencias

[1] “Automatic Personalization Based on Web Usage Mining”, Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, DePaul University, Chicago, University of Minnesota, Minneapolis.
[2] “Web Mining: Estado Actual de Invertigación”, Lic Gustavo D. Koblinc, Universidad de Buenos Aires.
[3] “Minería de Datos para análisis del uso de sitios web”, Luis Magdalena Layos, ETSI de Telecomunicación, Universidad Politécnica de Madrid
[4] “Introduction to Data Mining and Knowledge Discovery”, Third Edition, Two Crows Corporation.
[5] “Knowledge Discovery in Databases”, Project by Xue Wei Wang.
[6] “Fast Algorithms for Mining Association Rules”, Rakesh Agrawal, Ramakrishnan Srikant. IBM Almaden Research Center.